TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa

Título: “Análisis espacial de la relación entre mortalidad y vivir cerca de zonas verdes”

Autor: Emilio J Sánchez Díaz

Director: Dr. Xavier Basagaña Flores

Ponente: Dra. Mª Pilar Muñoz Gràcia

Centro: CREAL

Universidad: UPC - UB

Año académico: 2012/2013

UPC
Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA
UNIVERSITAT DE BARCELONA

Facultad de Matemáticas y Estadística
Universidad Politécnica de Cataluña
Universidad de Barcelona

Tesis de Máster

"Análisis espacial de la relación entre mortalidad y vivir cerca de zonas verdes"

Autor: Emilio J Sánchez Díaz

Director: Dr. Xavier Basagaña Flores

Ponente: Dr. M. Pilar Muñoz Gràcia

Quiero agradecer este trabajo al Dr. Xavier Basagaña por haberme ofrecido su ayuda en todo momento y haber tenido tanta paciencia, a la Dra. Mª Pilar Múñoz por ponerme en contacto con el CREAL, a Núria Almirall por ser una gran compañera y amiga y en especial a mis padres y hermana por estar siempre ahí.

Prefacio

Numerosos estudios han demostrado que existe una influencia positiva entre zonas verdes o naturales y las personas, sobretodo a lo referente a la salud en términos generales, en términos de salud mental, de años de vida, en términos de salud física, en la rapidez de recuperación, etc. Es por ello, que uno de los objetivos principales de este trabajo es observar cómo pueden influir las zonas verdes en la mortalidad de las personas. Ejemplo de estos estudios se encuentran, entre muchos otros, en los artículos de Maas et al (2006), *Green space, urbanity, and health: how strong is the relation?,* de Popham F. y Mitchell R. (2007), *Greenspace, urbanity and health: relationships in England,* y *Green space as a buffer between stressful life events and health,* de van den Berg A. E. et al (2010).

De confirmarse esta influencia en zonas metropolitanas de Barcelona se podrían ofrecer recomendaciones para formulaciones políticas y dar unas directrices profesionales a personas que se dediquen a la ordenación del territorio y de la salud para crear un medio ambiente sano. Es decir, este trabajo se enfoca desde una perspectiva ecológica que permita una mejor distribución del terreno, dadas las diferentes necesidades de las personas, entre ellas el gozar de una buena salud.

En consecuencia, esto puede tener impacto no solamente en el ámbito territorial, también en diferentes ámbitos económicos y sociales, como puede ser la generación de más o menos espacios para el uso de vehículos, así como un mayor número de zonas rurales o urbanas. Concienciando más a las personas de la importancia de preservar los parajes naturales, del posible impacto negativo que puede tener la expansión de zonas urbanas, de la influencia del cambio climático, la contaminación, de expandir más las zonas verdes como puede ser mediante la creación de parques, etc.

Hay que destacar la gran participación de muchas personas para obtener los datos. No solamente los pertenecientes a diferentes instituciones (CREAL e INE) si no a todos aquellos individuos que han aportado su opinión mediante encuestas.

Resumen

Introducción: A lo largo de muchos años se han realizado numerosos estudios de la influencia de factores externos en la mortalidad de la población. Entre estos factores, se han considerado algunos como el nivel económico, el estatus social…

Objetivo: En este proyecto se pretende hacer también el análisis de la mortalidad de la población en diferentes municipios del área metropolitana de Barcelona utilizando diversos factores e identificar *patrones espaciales* que justifiquen el comportamiento de la mortalidad (años 1999 a 2006) en término espaciales.

Métodos: Se trabaja con los datos por área censal y con la variable respuesta la *tasa de mortalidad estandarizada (SMR)*. Partimos de un análisis descriptivo de dichos datos, una vez realizado se hacen diferentes contrastes de hipótesis para detectar si existe una *autocorrelación espacial a nivel global y local* de la variable SMR. Por último se suaviza la SMR mediante un *modelo de heterogeneidad*, que no considera *efectos aleatorios espaciales* y un *modelo de Besag, York y Mollié (BYM)* que sí los considera.

Resultados: Se verifica que hay que considerar autocorrelación espacial de la variable SMR y también que la *suavización* es mejor con un modelo BYM. Además los factores en el modelo BYM que son significativos son el porcentaje de edificios con calefacción en el 2001, porcentaje de edificaciones vacías, el *índice de vegetación de diferencia normalizada (NDVI)* y la altura media de los edificios, la cual parece ser la única variable, que incluye el modelo ajustado, que reduce el *riesgo relativo*.

Conclusiones: Existe una autocorrelación espacial de la SMR y es preferible considerar para este estudio un modelo que tenga en cuenta los efectos espaciales aleatorios. Por otra parte, se confirma la necesidad de estratificar los datos, para tener resultados más realistas.

Palabras claves: *patrones espaciales, tasa de mortalidad estandarizada (SMR), autocorrelación espacial global, autocorrelación espacial local, efectos aleatorios espaciales, índice de vegetación de diferencia normalizada (NDVI) , riesgo relativo.*

MSC2000: 92B15

Abstract

Background: Over many years there have had numerous studies about the influence of external factors on mortality in the population. Among these factors are considered some as the economic, social status...

Objective: This project also aims to make the analysis of population mortality in different municipalities of the metropolitan area of Barcelona using various *spatial patterns* and identify factors that justify the behavior of mortality (1999 to 2006) in spatial terms.

Methods: It works with the data by census tract and with the response variable the *standardized mortality ratio (SMR*). We start with a descriptive analysis of the data, it followed by different contrasts hypothesis for detecting a *global and local spatial autocorrelation* of the SMR variable. Finally we smooth the SMR by a *heterogeneous model* that does not consider *spatial random effects* and also approach by a *Besag, York and Mollie (BYM) model* that considers its.

Results: It checks to consider spatial autocorrelation of the SMR and that smoothing is better with a BYM model. Besides the factors that are significant with a BYM model are the percentage of buildings with heat in 2001, percentage of empty buildings, the *normalized difference vegetation index (NDVI)* and the average height of the buildings, which seems to be the only variable, that includes the adjusted model, which reduces the *relative risk*.

Conclusions: It's confirmed the existence of spatial autocorrelation of the SMR and is preferable for this study consider a model that takes into account the random special effects. Moreover, it verifies the need to stratify the data, for more realistic results.

Keywords: *spatial patterns, standardized mortality ratio (SMR), global spatial autocorrelation, local spatial autocorrelation, heterogeneous model, Besag, York and Mollie (BYM) model, normalized difference vegetation index (NDVI), relative risk.*

MSC2000: 92B15

Notación

$\bar{x}$	Promedio de los valores de la variable x.
$\hat{x}$	Valor estimado de x.
E(X)	Esperanza de la variable aleatoria X.
V(X)	Varianza de la variable aleatoria X.
N(a,b)	Distribución normal de media a y varianza b.
$Poisson(a)$	Distribución de Poisson de parámetro a.
$loggamma\ (a,b)$	Distribución loggamma de parámetros a y b.

Índice

1. Introducción

A lo largo de muchos años se han realizado numerosos estudios de la influencia de factores externos en la mortalidad de la población. Entre estos factores, se han considerado algunos como el nivel económico, el estatus social… Como se puede ver en el artículo de *"Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain"* de *Maria Antòniva Barceló, Marc Saez y Carme Saurina*.

En este proyecto se pretende hacer también el análisis de la mortalidad de la población en diferentes municipios del área metropolitana de Barcelona, pero utilizando diversos factores externos entre ellos vivir cerca de zonas verdes, nivel educativo, económico... Además, otro punto a destacar es la identificación de patrones que justifiquen el comportamiento de la mortalidad en términos espaciales, es decir que dependiendo de donde se localice el lugar censal se obtengan unos resultados u otros.

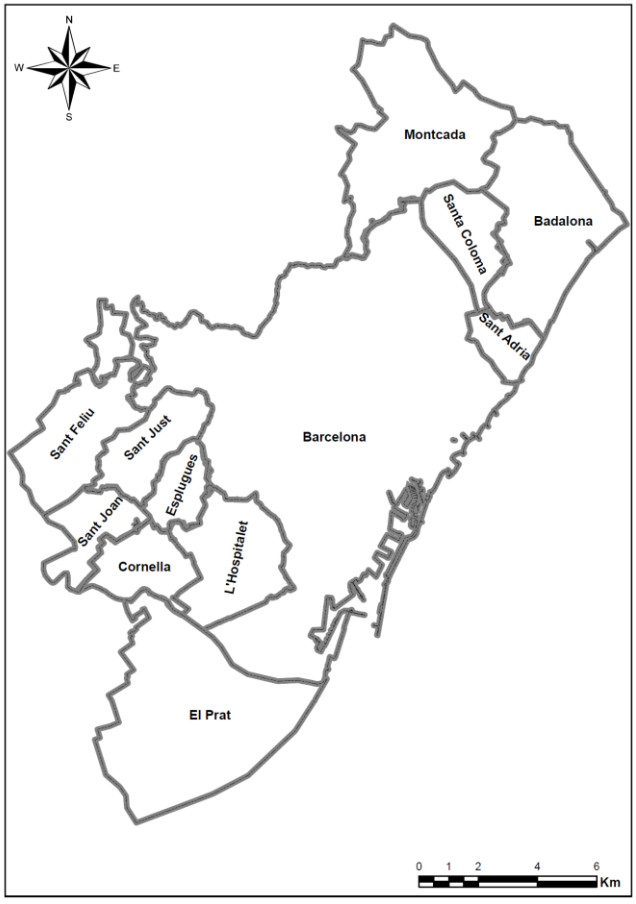


Figura1. Mapa de las zonas metropolitanas de Barcelona

La metodología empleada se basa en un análisis espacial, es decir se hará un estudio teniendo en cuenta la localización de los diferentes lugares donde estén censados los diferentes individuos de la población de estudio. Para ello, se destacan tres partes: la primera es un análisis descriptivo de los diferentes datos obtenidos, la segunda parte se hace un estudio de correlación espacial sobre la mortalidad entre los diferentes lugares de estudio y por último se realiza un modelo espacial que permita cuantificar la variación, suavizar las estimaciones de mortalidad de las áreas, y estudiar la influencia de los factores relacionados con la mortalidad como puede ser la presencia de zonas verdes.

Una vez obtenido los resultados, se analizarán y se llegará a unas conclusiones, que dependen de la metodología empleada y los resultados obtenidos.

2. Presentación de los datos

2.1 *Diseño*

La información de los diferentes lugares censales ha sido obtenida por el CREAL (*Centre de Recerca en Epidemiologia Ambiental*) de diferentes fuentes. Dicha información contiene la distribución de:

- La población por grupos de edad de cada zona censal del área metropolitana de Barcelona (2,203 áreas censales, obtenida del Instituto Nacional de Estadística).
- La mortalidad en cada área censal de Barcelona (2,203 áreas censales, datos obtenidos del registro de mortalidad de Cataluña, asignación a las secciones censales mediante el geocódigo de la dirección del fallecido).
- La mortalidad por grupo de edad en el área metropolitana de Barcelona (obtenida agregando la información anterior)
- Diferentes variables socioeconómicas, geológicas, geográficas y ambientales sobre diferentes zonas censales de Cataluña (2,203 zonas censales), obtenidas del Censo del 2001 (Instituto Nacional de Estadística).

Dicha información viene identificada mediante un número de varios dígitos (código censal), el cual me permite obtener el código de la provincia (08 si es Barcelona, 17 Gerona, 25 Lleida y 43 Tarragona), el código del municipio (cinco dígitos), el código del distrito (dos dígitos) y el código de la sección censal (tres dígitos).

Para obtener las coordenadas del lugar donde estaba censado el individuo se ha empleado el geo-código, un procedimiento por el cual sabiendo la dirección del individuo se pueden saber sus coordenadas *x* e *y,* de forma que su localización se pueda representar mediante unas coordenadas (latitud y longitud).

Las áreas metropolitanas de Barcelona estudiadas son Badalona, Barcelona, Cornellà del Llobregat, El Prat de Llobregat, Esplugues de Llobregat, Hospitalet de Llobregat, Montcada I Reixach, Sant Adrià del Besòs, Sant Feliu de Llobregat, Sant Joan Despí, Sant Just Desvern y Santa Coloma de Gramenet (*Figura1*).

2.2 Población de estudio

En este trabajo, la población llevada a estudio son los residentes en el área metropolitana de Barcelona, España, que fallecieron en el período desde 1999 hasta 2006.

Como se ha usado en otras investigaciones, se ha tomado el área censal como la unidad estadística más pequeña, ya que dicha unidad se ha reconocido como el área geográfica más óptima para el estudio de la variabilidad espacial de los niveles de salud en las ciudades. Además, el área censal es la unidad más pequeña de desagregación para el cual los datos socioeconómicos están disponibles.

En la *Tabla.A1* (apéndice final) se puede observar, para cada municipio, el número de áreas censales, la población total y la población por grupos de edad, así como la distribución de la población en las áreas censales. Donde en términos generales se obtiene una media de 1,086 individuos por lugar censal, siendo 7,003 individuos el máximo de individuos censados y 91 el mínimo, los valores de los cuartiles son 791 (primer cuartil) y 1,276 (tercer cuartil).

Algunos individuos han sido excluidos (2,385 individuos, es decir alrededor del 0.1% de la población total) debido a que no se pudo obtener bien su geo-código de su dirección, debido a que la dirección era incompleta, errónea, o porque el sistema de codificación no tenía esa dirección en la base de datos.

2.3 Variable respuesta

La variable respuesta utilizada es la tasa de mortalidad estandarizada (SMR) de cada sección censal como un indicador de la mortalidad. La SMR se define como la tasa entre individuos fallecidos observados e individuos fallecidos esperados, en base a la estructura de la sección censal y la tasa de mortalidad de referencia. Se ha tomado como referencia la población total de las áreas metropolitanas estudiadas. La fórmula para obtener dicha tasa es la siguiente:

$$SMR_i = \frac{\mathrm{M_i}}{\sum_{\mathrm{k=1}}^{\mathrm{I}} \mathrm{N_{ik}} \cdot \mathrm{MR_{ik}^*}}$$

donde,

- $\mathrm{M_i}$ = Número de individuos fallecidos de un lugar censal concreto.
- $\mathrm{N_{ik}}$ = Número de individuos del grupo k censados en un lugar concreto.
- $\mathrm{MR_{ik}^*}$ = Tasa de mortalidad bruta de la población de referencia (la población total). Es decir, el número de individuos muertos entre el número de individuos censados en al población total.
- I = número de agrupaciones de edad.
- i = Lugar censal del cual se calcula la SMR.

El uso de dicha variable es debido a la diferencia que hay entre las mortalidades en los diferentes grupos de edad y las diferentes estructuras de edad de las secciones censales. Además hay que destacar que no es necesario saber el número de individuos fallecidos de cada grupo de edad según el lugar censal de estudio y eso es un ahorro muy considerable.

Para tener dicho índice, no se han tenido en cuenta aquellos individuos fallecidos que están mal codificados en el geo-código ya que el sistema de geocodificación los asignaba en el centro de la ciudad. Su inclusión, hubiera provocado una SMR muy elevada en el centro de la ciudad. Por otra parte, las variables censales no se asignarían de forma correcta para esos sujetos. Por

otra parte, para algunos individuos el geo-código cae en la frontera entre dos áreas censales. Para esos casos, los individuos se asignaron a ambas pero con un peso de 0.5 para cada una de ellas.

Tabla1

SMR de cada municipio

	Área Metropolitana				
	Media	Std deviation	PCTL 25	PCTL 75	Mediana
Badalona	1.12	0.78	0.60	1.44	0.96
Barcelona	1.18	0.86	0.65	1.51	1.02
Cornellà de Llobregat	1.16	1.21	0.44	1.30	0.87
Esplugues de Llobregat	1.14	0.82	0.58	1.42	1.09
L'Hospitalet de Llobregat	1.23	1.11	0.59	1.57	1.00
Montcada i Reixach	1.11	0.80	0.62	1.08	0.82
El Prat de Llobregat	1.29	1.00	0.66	1.82	1.02
Sant Adrià del Besòs	1.77	2.28	0.57	1.91	1.17
Sant Feliu de Llobregat	1.04	1.22	0.17	1.36	0.68
Sant Joan Despí	1.28	0.97	0.70	1.51	0.99
Sant Just Desvern	1.16	0.51	0.84	1.45	1.23
Santa Coloma de Gramenet	1.30	0.87	0.75	1.50	1.17
Total	1.19	0.92	0.63	1.51	1.02

Se puede observar que la media de todos los individuos es mayor que uno (*Tabla 1*), lo cual significa que el número de individuos fallecidos observados es mayor del esperado, pero esto es debido a que algunos áreas censales presentan valores muy altos o muy bajos, las cuales son muy influyentes para la media del municipio, por lo tanto es mejor hacer la interpretación mediante la mediana.

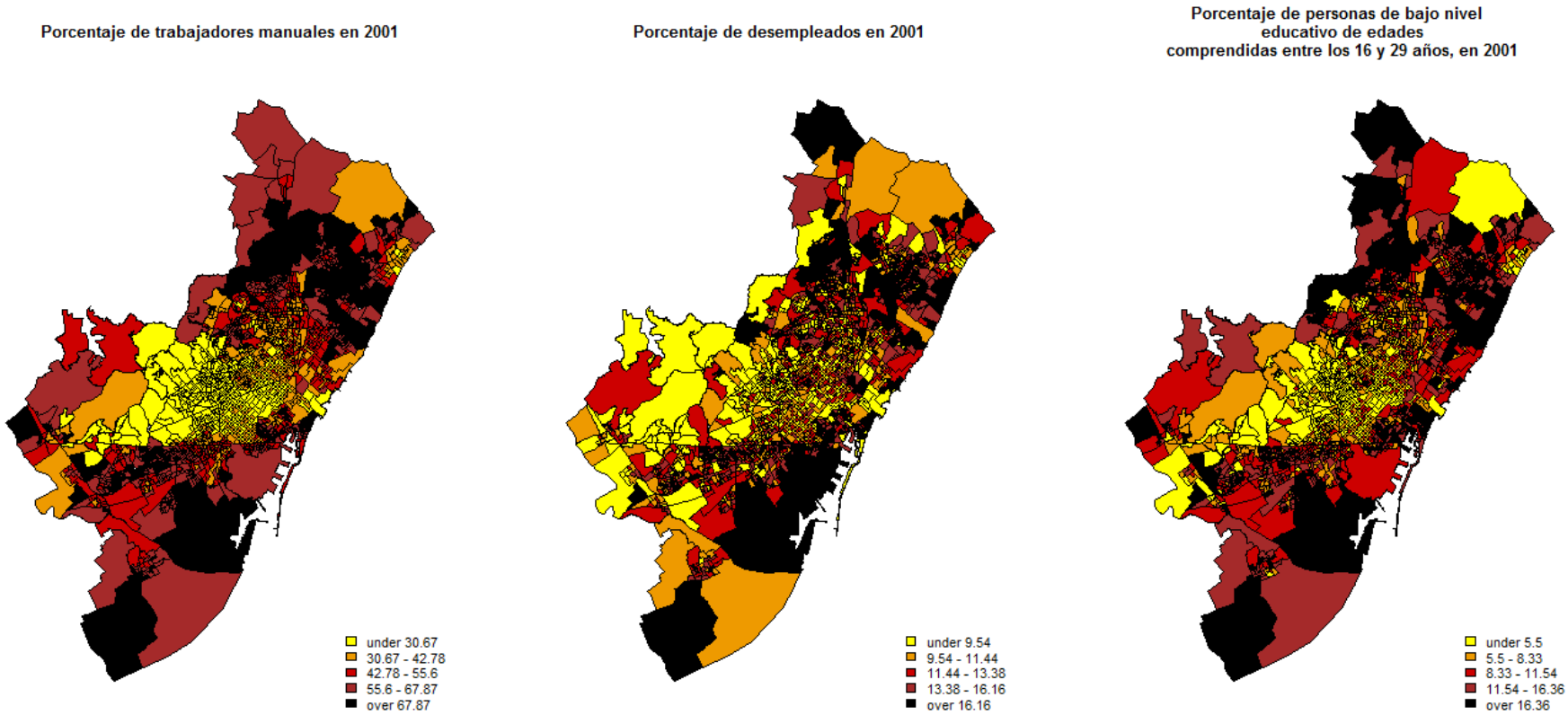
El municipio de Sant Feliu de Llobregat presenta una SMR mediana del 0.68, comparado con el resto es el índice más bajo. En cambio, St Just Desvern es el que presenta una SMR mediana del 1.23, la más elevada.

2.4 *Variables socieconómicas*

Las tres variables socioeconómicas incluidas en el estudio pertenecen a las diferentes zonas censales de Cataluña en el 2001. Se toma los valores del 2001 como un valor promedio de los años de estudio, hay que tener en cuenta que muchos datos obtenidos vienen dados por un estudio censal, y al ser un trabajo tan costoso se suelen recoger datos cada 10 años. Puesto que los

Figura2

Mapas por cuantiles que representan la distribución socioeconómica de las diferentes zonas censales, según el empleo y la educación



datos del 2011 son muy recientes para tenerlos todavía, se trabajan con los del 2001. Se espera que aunque los valores de dichas variables hayan variado durante los años de estudio (1999-2006), su distribución espacial en 2001 sea representativa de la distribución espacial durante todo el periodo. Dichas variables son:

- Trabajadores manuales: porcentaje de trabajadores que por lo menos tengan 16 años de edad y que se dediquen a oficios como empleados de un restaurante, trabajar en servicios personales, trabajar como vigilantes o dependientes en una tienda; trabajar en oficios relacionados con la pesca o la ganadería; aquellos que sean artesanos, empleados que estén en la parte de manufacturación de una fábrica, que se dediquen a la minería o a la construcción industrial, pero no se incluyen instaladores, operadores mecánicos y ensambladores que operen en una fábrica; aquellos trabajadores no calificados también son incluidos en este grupo.
- Desocupados: porcentaje de individuos, a partir de 16 años, sin empleo o activamente buscando empleo, con respecto el total de individuos económicamente activos.
- Nivel de educación: porcentaje de individuos mayores o con 16 años de edad, con menos de cinco años de escolarización o con más de cinco años de escolarización, pero sin completar su educación básica. Dicho porcentaje se calcula respecto el número de personas que como mínimo tengan 16 años de edad.

Se puede observar en la *Tabla.A2* (apéndice final) la información sobre las variables socioeconómicas utilizadas en este estudio. En el mapa *Figura2*, se observa cómo se distribuyen estos porcentajes. Así donde menos trabajadores manuales hay parece que es en el centro de Barcelona y la zona del oeste, lo mismo ocurre con el nivel de educación, estas zonas son las que tienen un menor número de individuos que no hayan completado la educación básica. Es decir, la zona oeste y central de Barcelona, se puede categorizar como las zonas con mayores valores socioeconómicos.

2.5 *Zonas verdes y contaminación*

Se distinguen dos tipos de datos ambientales, que se pueden clasificar en subjetivos u objetivos. Los datos subjetivos sobre zonas verdes provienen

de las respuestas de los individuos al cuestionario del censo. En concreto, la pregunta del censo era "¿Tiene su vivienda pocas zonas verdes? (jardines…)" (ver *Encuesta* en apéndice final). Se dispone del porcentaje de habitantes de cada sección censal que contestaron afirmativamente a esta pregunta.

Por otra parte tenemos los datos de contaminación y el *índice de vegetación de diferencia normalizada (NVDI)* del año 2002. El NVDI es un índice usado para estimar la cantidad, calidad y desarrollo de la vegetación con base a la medición, por medio de sensores remotos instalados comúnmente desde una plataforma espacial, de la intensidad de la radiación de ciertas bandas del espectro electromagnético que la vegetación emite o refleja. Así, la vegetación aparece reflejada relativamente oscura en la región de radiación fotosintética activa y relativamente brillante en el infrarrojo cercano. En contraste, las nubes y la nieve tienden a ser bastante brillantes en el rojo así como también en otras longitudes de onda visibles, y bastante oscura en el infrarrojo cercano.

El índice de vegetación de diferencia normalizada, NDVI, se calcula a partir de estas medidas individuales de la siguiente manera:

$$NDVI = \frac{(IRCercano - ROJO)}{(IRCercano + ROJO)}$$

En donde las variables ROJO e IRCercano están definidas por las medidas de reflexión espectral adquiridas en las regiones del rojo e infrarrojo cercano, respectivamente. Estas reflexiones espectrales son en sí cocientes de la radiación reflejada sobre la radiación entrante en cada banda espectral individual; por tanto, éstos toman valores entre un rango de 0.0 a 1.0. El NDVI varía como consecuencia entre -1.0 (menos vegetación) y +1.0 (más vegetación). La *Figura3a* muestra los valores de estas variables, siendo visible como las zonas claras del mapa que representa la distribución de la variable poco verde, son zonas oscuras en el mapa que refleja la distribución de la variable NVDI, es decir no se contradicen lo único que su interpretación es diferente. No obstante, hay que destacar que entre ellas su correlación es del -0.37, esto significa que la opinión de las personas y los valores obtenidos del NDVI son dependientes, esto es debido a que las variables no muestran la misma, el índice NVDI muestra en el mapa mucha vegetación en

Figura3a
Mapas por cuantiles que representan la distribución de las variables medioambientales

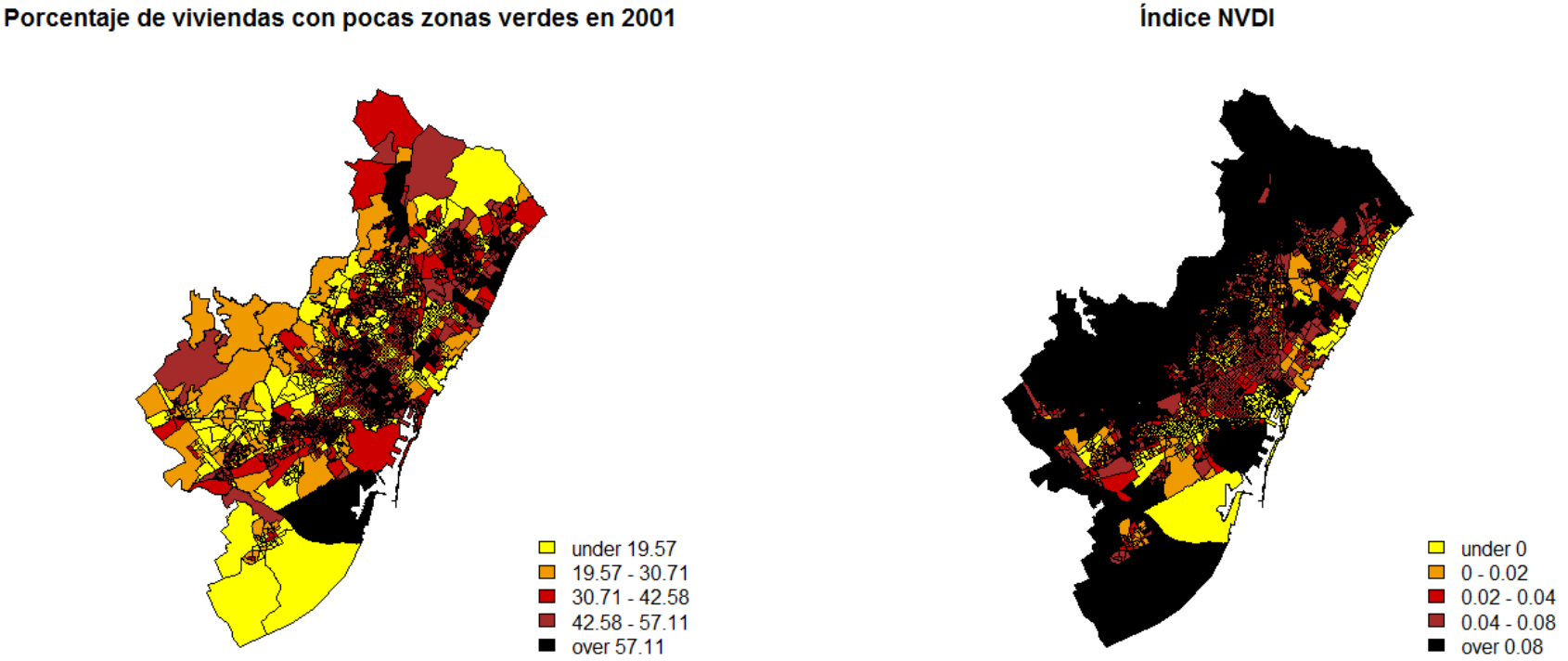
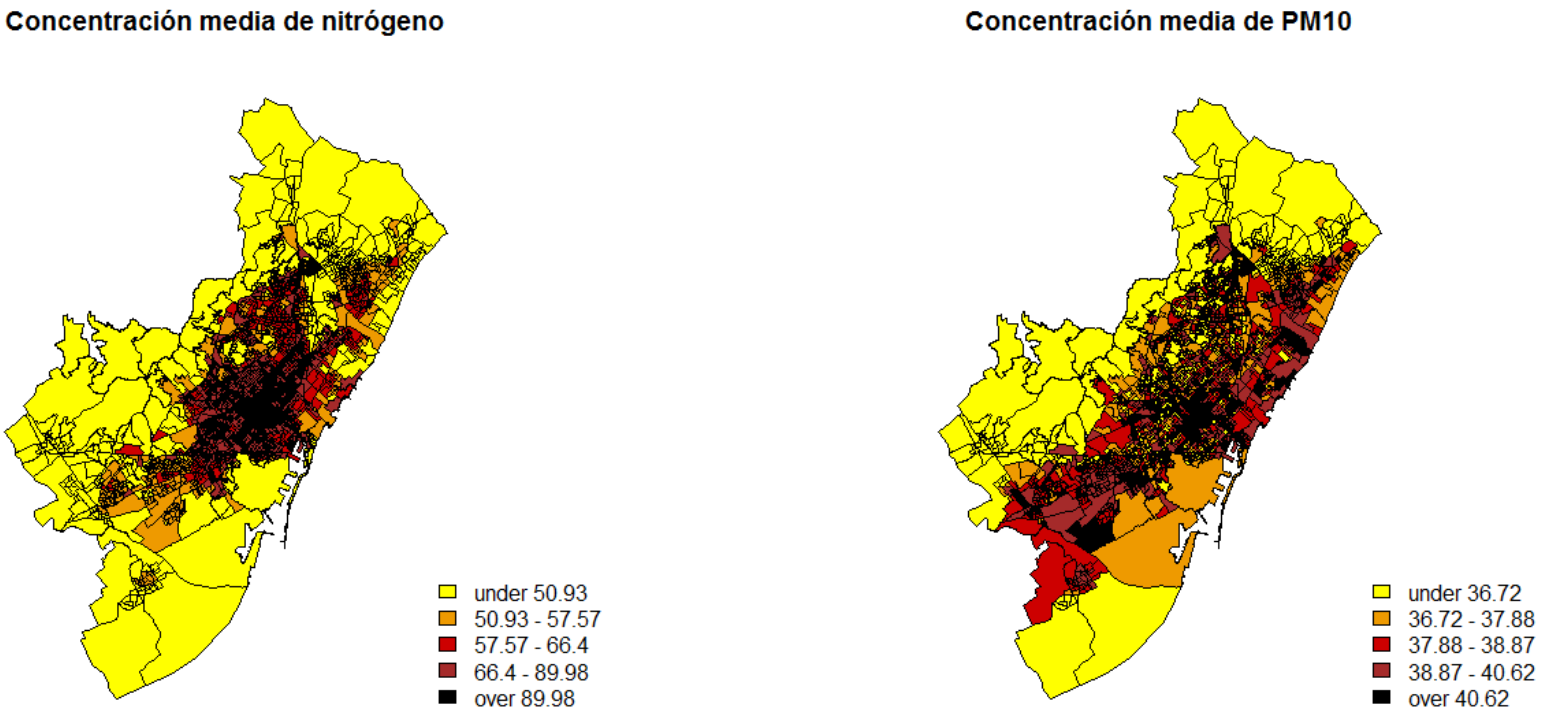


Figura3b *(continuación)*



la zona de Collserola, mientras que la variable referente a la encuesta no, ya que hace referencia más a parques y jardines que no a bosque.

De los datos de contaminación se dispone de la concentración media de dióxido de nitrógeno y de PM10 (pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro es menor que 10 μm). Estos datos han sido obtenidos mediante medidas ambientales de los contaminantes en diferentes puntos, bien caracterizados con respecto a variables como el tipo de calle, densidad de tráfico, altura de edificios, etc., combinados con una modelización mediante modelos *LUR* (*land-use regressions*). Estos modelos se basan en el principio de que las concentraciones de contaminantes en cualquier ubicación dependerán de las características físicas y ambientales de la zona circundante particularmente aquellos que influyen o reflejar la intensidad de emisión y eficiencia de dispersión. El modelo se realiza mediante la construcción de las ecuaciones de regresión múltiple que describen la relación entre las concentraciones medidas en una muestra de sitios estudiados mediante herramientas tecnológicas, y a las variables ambientales calculadas, utilizando *GIS* (*Sistema de Información Geográfica*, es una herramienta y base de datos que ayuda al estudio de problemas de gestión y planificación geométrica), para las zonas de influencia alrededor de cada sitio. La ecuación resultante se utiliza entonces para predecir las concentraciones en lugares no medidos sobre la base de estas variables de predicción. La predicción puede hacerse, sobre localizaciones de puntos específicos (por ejemplo, direcciones residenciales) o para una malla fina, etc. Al final con la intersección con los datos de población a nivel de área y un mapa se puede estimar la distribución de la exposición. Los datos de contaminación son un input de este proyecto, su modelización no fue parte de este trabajo.

En la *Figura3b* se puede observar la distribución espacial de los contaminantes, se puede ver como los niveles de dióxido de nitrógeno de la zona de Barcelona están muy por encima del resto de zonas, sobretodo en la parte central. Por otra parte parece que hay una concentración alta de pequeñas partículas en l'Hospitalet de Llobregat, Cornellà, el norte del Prat y en general gran parte de Barcelona. La correlación entre estas variables (concentración de dióxido de

nitrógeno y de pequeñas partículas) también es baja, alrededor del 0.47, además las correlaciones de estas variables con las variables poco verde y NDVI, son muy cercanas a cero, puesto que el caso donde se da mayor correlación es entre la variable NDVI y las partículas minúsculas con un valor del -0.21.

2.6 Tipos de vivienda

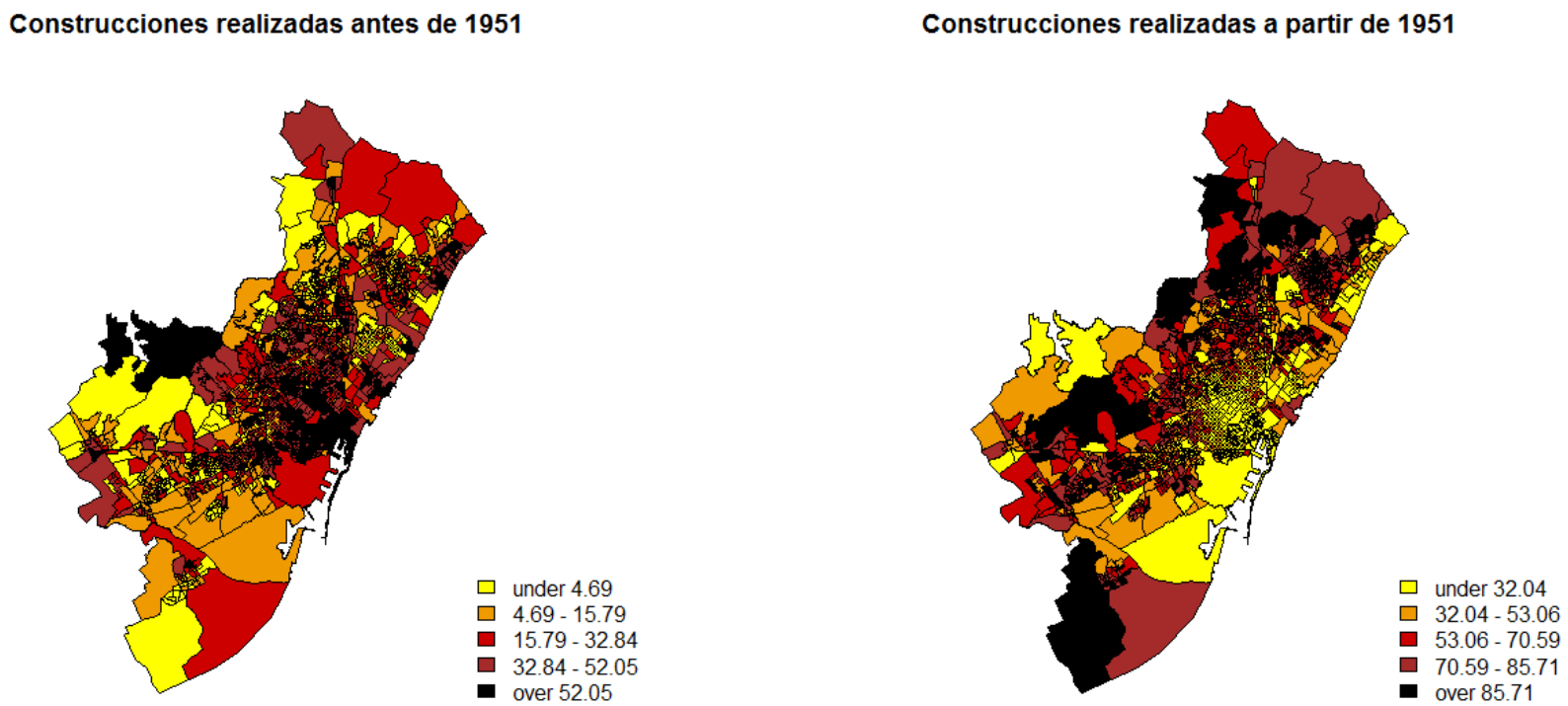
En este estudio otros datos que se han tenido en cuenta son los distintos tipos de vivienda que tienen los habitantes de cada zona censal. Este tipo de variable puede jugar el papel de variable socioeconómica, es como un tipo de medida que indica el estatus social que se localiza en el área censal. Los datos del tipo de vivienda que se disponen son los siguientes:

- Primeramente, cuándo los hogares fueron construidos y también el número de hogares vacíos (en la *Figura.A1* del apéndice final se puede observar dicha distribución por ciudad). El comportamiento de dichas variables es el mismo en todas las áreas metropolitanas estudiadas, en general, se puede observar que entre 1961 y 1970 muchas construcciones fueron realizadas en este intervalo de tiempo, quizás porque mucha gente emigró a la provincia de Barcelona en busca de un trabajo.

  En este caso, se propuso la siguiente agrupación de los datos, según si la construcción de las edificaciones fue realizada a partir de 1951 o antes, se ha decidido hacerlo así ya que a partir de 1951 los porcentajes del número de construcciones según la época tienden a ser significativamente más grandes que los valores anteriores (observad *Figura.A1*). Para verificar este agrupamiento se ha realizado un contraste de Kruskal-Wallis, de esa forma se puede considerar que los datos representan lo mismo, es decir, que no hay diferencias entre las épocas. Los resultados obtenidos en la *Figura4* muestran como a partir del 1951 en la zona del litoral no ha habido muchas construcciones, cosa que en el mapa que indica el porcentaje de construcciones previas al año 1951 muestra lo contrario, es normal, ya que muchas zonas que han sido edificadas con anterioridad después no siguen siendo edificadas por falta de terreno, y lo mismo ocurre a la inversa, es decir, zonas que no han

Figura4

Mapas por cuantiles que representan la distribución del porcentaje de edificaciones construidas anteriormente y posteriormente del año 1951.



sido edificadas previamente al año 1951, suelen ser los lugares con más construcciones a partir del 1951 debido a la cantidad de terreno no explotado.

- Otra clasificación es según el porcentaje de hogares con calefacción en 1991, con refrigeración en 1991 y también con calefacción en el 2001 y con refrigeración en el 2001. La *Tabla5a* y la *Tabla5b* muestran los resultados, los cuales nos permiten observar las diferencias entre 1991 y 2001, en 2001 muchos hogares tienen calefacción. Por otra parte se observa como la refrigeración en el hogar siempre ha sido menos habitual que la calefacción.
- Tercero, según la altura media de los edificios, medida en número de plantas. Si se observa la *Figura5,* se puede ver como la zona de Barcelona y la zona oeste de El Prat son los lugares donde las edificaciones tienen una altura media más elevada, la altura de los edificios depende mucho de la densidad de población, por lo tanto es natural haber obtenido estos resultados.

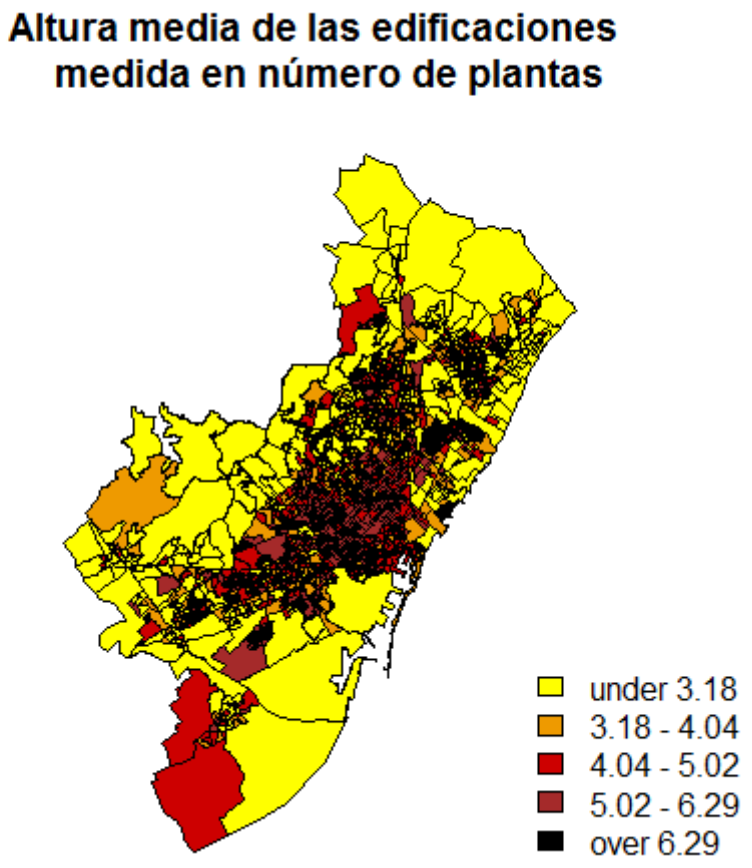


Figura5. Distribución por cuantiles de la altura media de las edificaciones medida en número de plantas.

Tabla 5a
Información sobre el porcentaje de viviendas con calefacción y refrigeración en el 1991

	Viviendas con calefacción %				Viviendas con refrigeración %			
	Mean	Std deviation	PCTL 25	PCTL 75	Mean	Std deviation	PCTL 25	PCTL 75
Badalona	13	13.8	4.9	18.1	4.8	3.1	3.2	5.7
Barcelona	29	23.9	10.3	41.9	6	4.9	3.7	7
Cornellà de Llobregat	11	8.4	6.5	11.5	5.1	1.7	4	5.9
Esplugues de Llobregat	31.6	24.1	14	43.7	6.8	2.1	5.4	7.7
L'Hospitalet de Llobregat	8.6	9	4.1	9.2	4	1.7	2.9	5
Montcada i Reixach	25.1	14.9	12.9	34.6	6.7	1.5	5.7	7.8
El Prat de Llobregat	13.2	16.8	4.3	13.2	4.5	2.3	3	6.2
Sant Adrià del Besòs	13.5	21.5	2.9	11.9	3.4	1.7	2.2	4.7
Sant Feliu de Llobregat	13.4	10.8	5.5	17.9	5.1	1.5	4	6.4
Sant Joan Despí	17.8	26.6	3.8	16.4	5.5	2.2	3.8	6.7
Sant Just Desvern	44.2	24	29.7	50.3	6.4	1.5	5.4	7.3
Santa Coloma de Gramenet	8.2	7.1	3.7	9.5	4.4	2.6	3.1	5
Total	23.6	2.3	6.9	33.9	5.5	4.3	3.5	6.5

Tabla 5b
Información sobre el porcentaje de viviendas con calefacción y refrigeración en el 2001

	Viviendas con calefacción %				Viviendas con refrigeración %			
	Mean	Std deviation	PCTL 25	PCTL 75	Mean	Std deviation	PCTL 25	PCTL 75
Badalona	37.5	16.4	25	50.8	14.7	6.1	10.4	17.5
Barcelona	49	18.6	35.2	60.3	20.6	9.1	14.3	25.8
Cornellà de Llobregat	36.7	14	25.6	43.8	19.2	5.4	16.9	21.7
Esplugues de Llobregat	50.2	19.8	36.3	64.8	20.9	8.2	15.7	22
L'Hospitalet de Llobregat	29.1	10.1	21.7	33.8	16.7	5.7	12.8	20
Montcada i Reixach	64	11.5	59.4	72	14.6	5.1	11.2	17.8
El Prat de Llobregat	36.1	10	27.5	43.4	19.8	5.7	17.2	22.8
Sant Adrià del Besòs	35.1	19.1	19.2	47.7	12.8	8	5.1	18.9
Sant Feliu de Llobregat	46.6	20.8	28.1	59	21.1	6.2	16.6	25.2
Sant Joan Despí	51.5	25.3	30.5	70.9	24.4	6.2	20	26.2
Sant Just Desvern	70.4	16.3	67.4	75.9	24.4	6.7	21.4	25.4
Santa Coloma de Gramenet	30.8	9.7	29.3	36.8	14.7	4.6	11.5	17
Total	44.8	18.9	29.9	56.2	19.4	8.5	13.4	24.4

Figura6
Mapas por cuantiles que representan la distribución del porcentaje de viviendas unifamiliares, hogares que son solamente viviendas, que sean viviendas y locales y por último que sean solamente locales

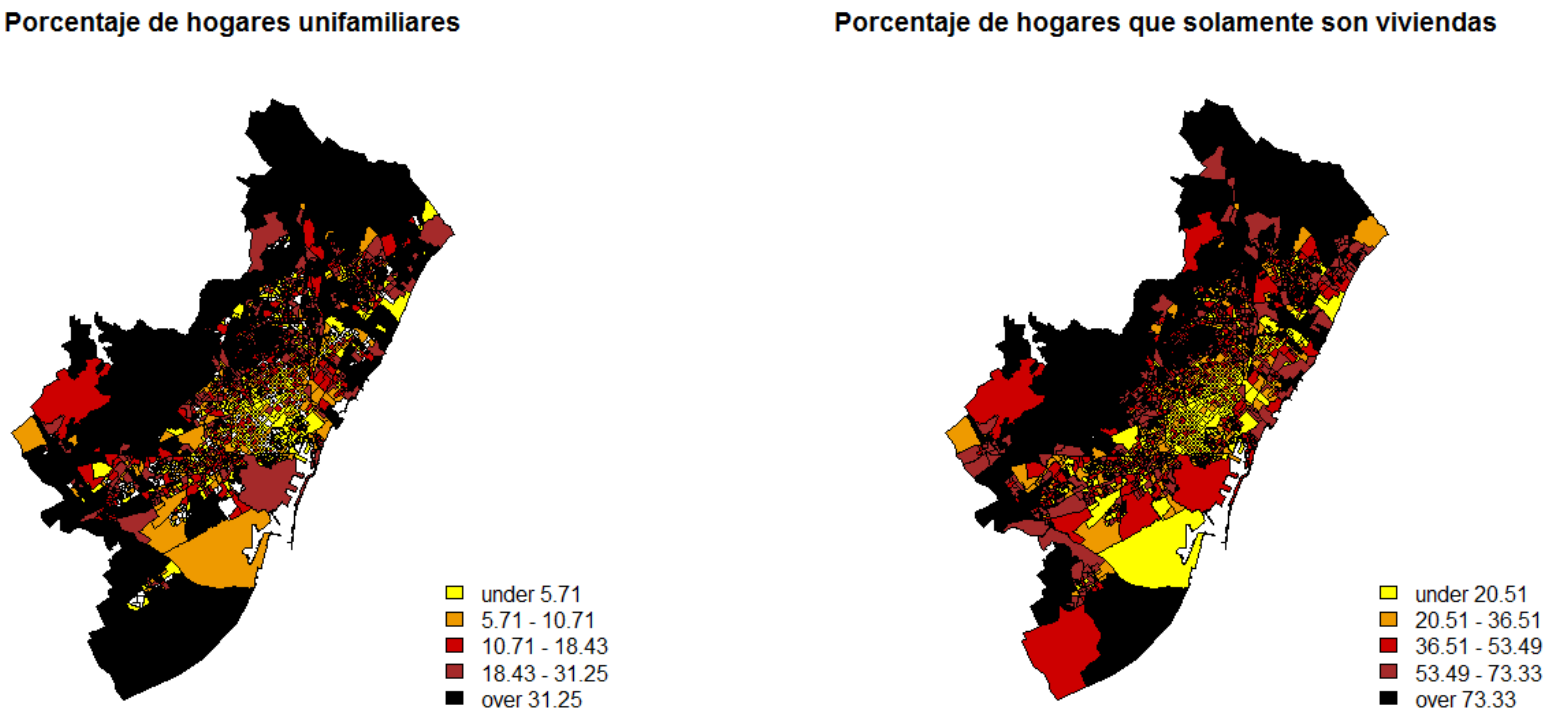
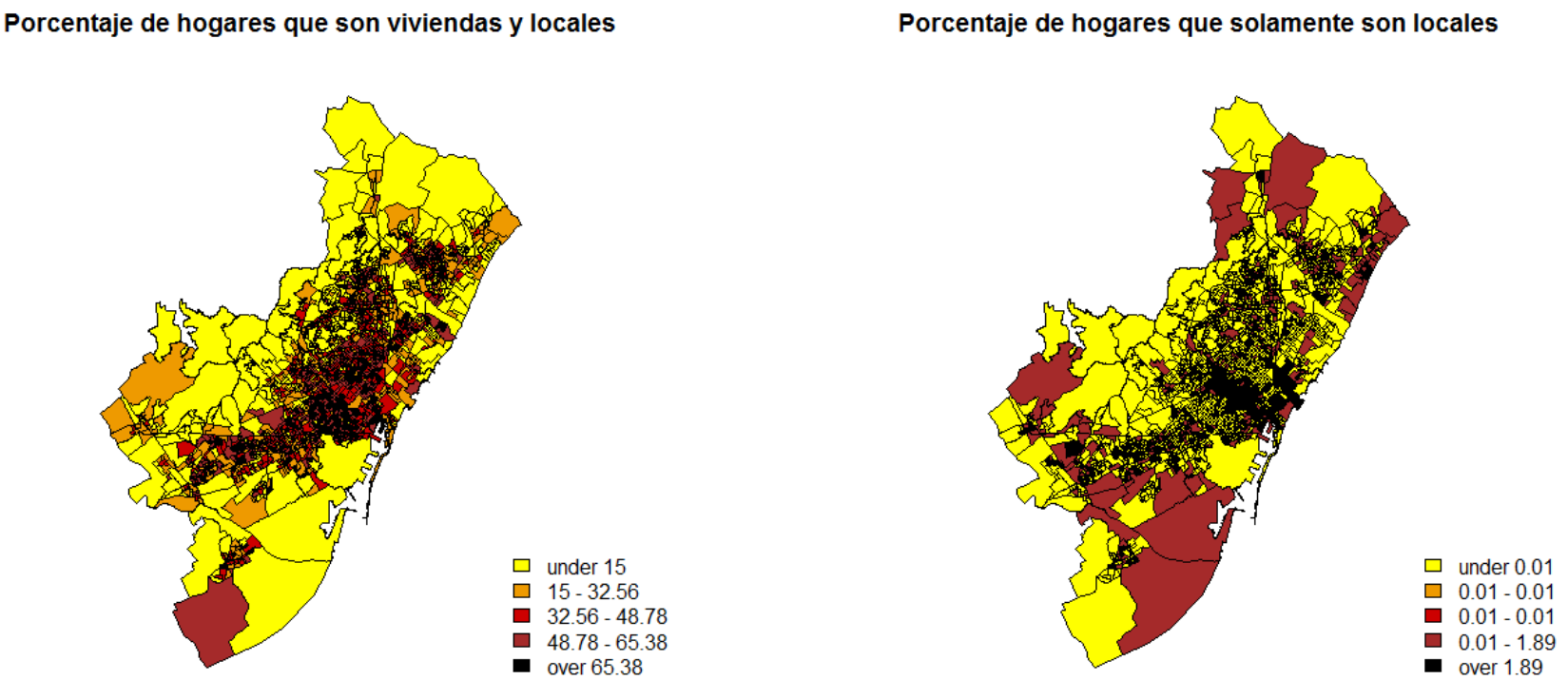


Figura6 *(continuación)*



- Finalmente, según el tipo de vivienda, es decir, si el hogar es unifamiliar; si el hogar es solamente vivienda, este caso incluye las viviendas unifamiliares; si el hogar es vivienda y local y por último si en vez de un hogar es un local. Cabe destacar como en este caso la variable vivienda y la variable vivienda y local tienen correlación negativa (con un valor según la correlación de Pearson de -0.89), por otro lado el número de hogares que solamente son locales tienen un nivel bajo en todos los municipios. La *Figura6* muestra como los porcentajes más elevados de viviendas unifamiliares y de hogares que sean solamente viviendas se sitúan por Sant Just Desvern, norte de Barcelona, por El Prat de Llobregat y Montcada i Reixach. También se puede observar que en general el número de hogares que sean solamente locales es muy bajo y que sean viviendas y locales también, aunque por la zona sur de El Prat de Llobregat y el centro de Barcelona es más elevado.

Las viviendas y las construcciones son elementos que aportan mucha información sobre las secciones censales y su situación actual, es por ello que se las considera como un tipo de variables interesantes de estudiar para incluir en el modelo predictivo.

3. Análisis exploratorio espacial

En esta sección se explorará si las razones de mortalidad estandarizadas (SMR) de las secciones del área de Barcelona presentan un patrón espacial, de confirmarse dicho patrón es razonable pensar que un posible buen modelo para el estudio de la SMR sea un modelo de autocorrelación espacial.

Se entiende como patrón espacial de la variable SMR, a la detección de un comportamiento en la distribución de dicha variable en el área estudiada, es decir verificar que existe una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar.

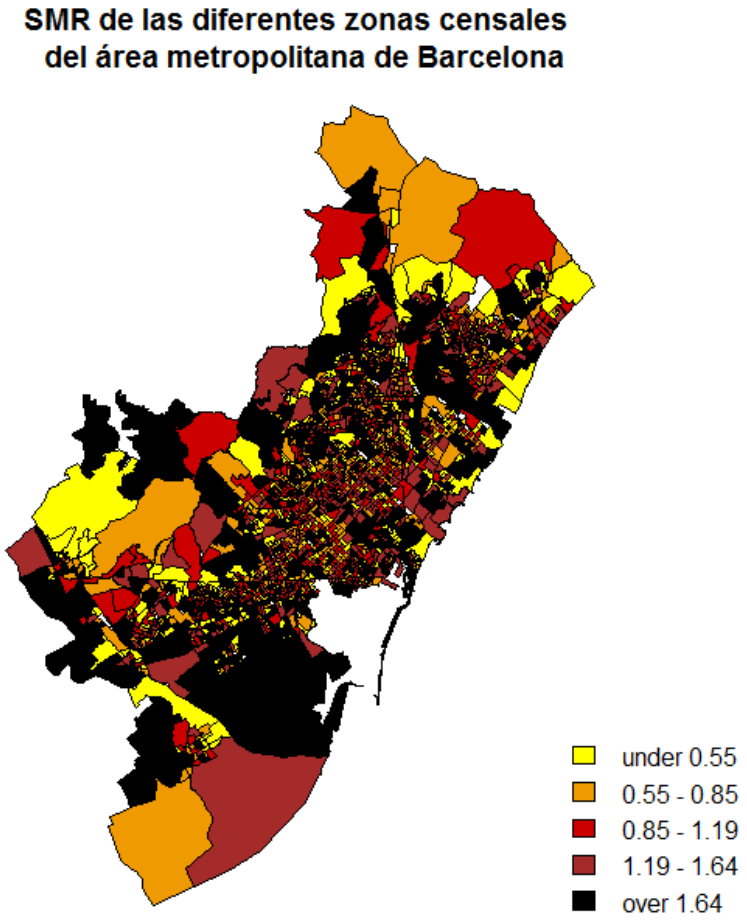


Figura7. Mapa de las SMR de las zonas censales del área metropolitana de Barcelona

Para contrastar si existe un patrón espacial, se tomará como base exploraciones de gráficas (detectando o no similitud de valores) y varios test basados en el contraste de hipótesis de la existencia de correlación espacial a nivel local y a nivel global.

3.1 Método de retículas o "lattice"

El método empleado para el estudio espacial de la SMR es el llamado método de retículas o más conocido como "lattice". Dicho método consta de una muestra de n datos de interés **Z(s)**, donde la componente **s** ϵ D (espacio donde se localizan las retículas). Para cada celda de la retícula, indexada por su localización **s**, se observa un solo valor de la variable **Z**. En los datos del proyecto se tiene:

- D es el dominio espacial o área de interés (el área metropolitana de Barcelona).
- $\mathbf{Z}(\mathbf{s})=(z_1(\mathbf{s}),\ldots,z_n(\mathbf{s}))$ datos de interés (SMR).
- $\mathbf{s}=(\mathbf{s}_{1,\ldots,}\mathbf{s}_n)$ hace referencia a las localizaciones de las secciones censales. Normalmente, se trabaja con dos coordenadas y tenemos $\mathbf{s}_i=(x_i,y_i)$. En este caso, corresponde a las coordenadas geográficas de los diferentes lugares censales estudiados.

No hay que confundir el método lattice, con el método geoestadístico general o el de patrones puntuales. Todos ellos se agrupan dentro del análisis espacial, pero son diferentes:

- En el método geoestadístico espacial los eventos de interés varían de forma continua sobre un dominio fijo, pero solo se observan en puntos (discretos), a diferencia del lattice, donde se observa un solo dato representativo de todos los puntos de la retícula, normalmente información agregada.
- En patrones puntuales se observan solo las coordenadas de los puntos donde ocurren eventos. Las coordenadas geográficas son pues la variable aleatoria de interés. En contraposición, en lattice las coordenadas de las celdas son fijas y conocidas, la aleatoriedad reside en el valor de la variable respuesta para cada celda.

3.2 Matriz de vecindad

En el espacio las relaciones son multidireccionales, es por ello que se requiere de una notación específica: **W**, conocida como matriz de pesos espaciales y más de forma informal matriz de vecindad. La matriz de pesos permite calcular el llamado *operador de retardo espacial*. Dicho operador se obtiene como el producto de la matriz de pesos espaciales por el vector de observaciones de una variable aleatoria, en este caso la SMR. Cada elemento del operador del retardo espacial es el promedio ponderado de los valores de la variable SMR en el subgrupo de observaciones vecinas.

El operador **W** es una matriz de *n* filas y *n* columnas definidas positivas, con ceros en la diagonal, y en principio, simétrica:

$$W = \begin{bmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & 0 & \dots & w_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & 0 \end{bmatrix}$$

Lo primero que se tiene que tener en cuenta al crear dicha matriz es qué consideramos vecino. En el caso particular de este estudio se ha considerado zona censal vecina aquellas zonas que compartan frontera con la zona de referencia, este criterio se le llama tipo *"queen"* (por el movimiento de la reina del juego del ajedrez) y es de primer orden, es decir, solamente se considera vecinos aquellos que comparten frontera con mi área de referencia directamente. Aquellas zonas que no son vecinas tienen un peso igual a cero:

$$w_{ij} = \begin{cases} 1 & si\ la\ región\ i\ y\ j\ comparten\ frontera \\ 0 & en\ caso\ contrario \end{cases}$$

Hay otras formas de considerar vecinos, por ejemplo, vecinos de tipo *"torre"* pero dicho tipo se usa más cuando se trabaja sobre una zona espacial en forma de cuadrícula. Se pueden dar casos de vecinos de segundo orden, es decir, cuando se consideran también vecinos, aquellas áreas que son vecinas de los vecinos de primer orden e incluso se pueden considerar criterios propios para definir qué se considera vecino.

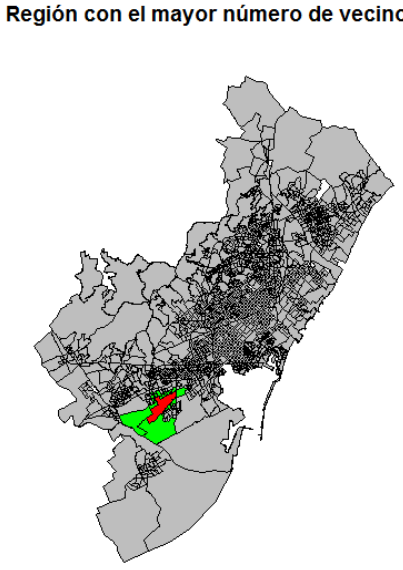


Figura8. Región con 25 vecinos, usando la matriz tipo queen.

La matriz de vecindad utiliza unos pesos w_{ij}, que suelen tomar diferentes valores según los propios intereses. Normalmente, en la mayoría de estudios, se suelen tomar los pesos estandarizados por filas, esto sucede cuando no se quiere dar más peso a unos vecinos que a otros, así la variable retardada espacialmente representaría los valores vecinos suavizados, dado que la suma de todos los pesos de una determinada fila es igual a 1**.** En el estudio, también se ha utilizado este criterio, el cual es el más utilizado comúnmente, es decir la suma de cada fila de la matriz es igual a uno, teniendo cada peso de una

misma fila el mismo valor. Existen muchos otros, pero que son más usados para casos particulares: los pesos binarios (si son vecinos imponemos que el peso tenga valor uno y si no valor igual a cero), *Anselin* (usa como peso la inversa de la distancia euclídea entre la regiones), *Cliff Odr*…

3.3 Autocorrelación espacial a nivel global

Para ver si existe un patrón de comportamiento espacial en los datos se computan diferentes estadísticos. En nuestro estudio de autocorrelación espacial de la SMR se han utilizado tres herramientas: el *contraste de la I ley de Moran, el Moran Scatterplot y el contraste de la G de Getis y Ord*.

- *Contraste de la I ley de Moran (Moran, 1948)*

  La I ley de Moran parte de la hipótesis nula de ausencia de autocorrelación espacial. Con la ayuda de la siguiente fórmula:

$$I = \frac{N}{S_0} \frac{\sum_{ij}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} i \neq j$$

  donde

  S_o = es el sumatorio de todos los pesos de la matriz.
  N = número de observaciones.
  x = es la variable de interés, en este caso la SMR.
  w_{ij} = es el peso correspondiente de la matriz **W,** fila *i* y columa *j*.

  Mediante la fórmula se puede detectar que este estadístico se parece mucho al cálculo de un coeficiente de correlación, pero añadiendo el factor espacial mediante los pesos.

  La forma de interpretar Z(I), el contraste de la I Ley de Moran, es la siguiente:

$$Z(I) = \frac{I - E(I)}{\sqrt{V(I)}}$$

  si Z(I) no es estadísticamente significativo se concluye que no existe autocorrelación espacial. Si es estadísticamente significativo y con signo positivo se tiene una autocorrelación positiva (predominio de concentración de valores similares en regiones vecinas), pero si es

significativo y con signo negativo, se tiene autocorrelación negativa (predominio de concentración de valores disímiles en regiones vecinas).

La esperanza de I, E(I), y la varianza de I, V(I) bajo la hipótesis nula de independencia espacial:

$$E(I) = \frac{-1}{N-1}$$

$$E(I^2) = \frac{N\left[(N^2-3N+3)S_1 - NS_2 + 3S_0^2\right] - b_2\left[(N^2-N)S_1 - 2NS_2 + 6S_0^2\right]}{(N-1)(N-2)(N-3)S_0^2}$$

Donde,

$$b_2 = \frac{m_4}{m_2^2}, \quad m_2 = \frac{\sum_i (x_i - \bar{x})^2}{N}, \quad m_4 = \frac{\sum_i (x_i - \bar{x})^4}{N} \qquad w_{i.} = \sum_{j=1}^{N} w_{ij}$$

$$S_1 = \frac{1}{2}\sum_i \sum_j (w_{ij} + w_{ji})^2 \qquad S_2 = \sum_i \sum_j (w_{i.} + w_{.i})^2$$

Z(I) sigue una distribución asintótica N(0,1), para una muestra grande, la cual puede utilizarse para la inferencia.

Otras muchas veces se suelen calcular unos pseudo-niveles de significación a partir de una distribución empírica derivada siguiendo un criterio de aleatoriedad condicional o de permutación.

En la siguiente tabla se pueden observar los resultados obtenidos del contraste de la I ley de Moran realizado sobre la variable SMR:

Matriz tipo queen				
Tipos de Pesos	**Estadístico I**	**E(I)**	**V(I)**	**p-valor**
Sin estandarizar	0.0434840696	-0.0004657662	0.0001453572	0.0001335
Estandarizados por filas	0.0234831839	-0.0004657662	0.0001533323	0.02655

Se observa cómo se confirma una autocorrelación espacial a nivel global ya que se rechaza la hipótesis nula. No obstante, se hicieron más pruebas con otros tipos de pesos. En concreto, se consideraron los puntos centrales de cada zona y se midió la distancia entre sus regiones vecinas, de tal forma que aquellas zonas más próximas tuvieran mayor peso. Los resultados fueron los siguientes:

Matriz donde dependiente de la distancia				
Tipos de Pesos	**Estadístico I**	**E(I)**	**V(I)**	**p-valor**
Sin estandarizar	0.0153261710	-0.0004657662	0.0001820299	0.1209
Estandarizados por filas	0.0280915823	-0.0004657662	0.0001749575	0.01543

En este caso se observa cómo dependiendo de los pesos utilizados se acepta la hipótesis nula o no, así si se utilizan unos pesos que dependan de la distancia a los centroides de cada zona sin estandarizar, no se considera que la variable SMR tenga una autocorrelación espacial. No obstante, el hecho de haber detectado correlación espacial con varias de las métricas nos indica que realmente existe algún tipo de patrón espacial.

- *Contraste de G de Getis y Ord (Getis and Ord, 1992)*

Las medidas de autocorrelación espacial se ven afectadas por el denominado *problema de la unidad de área modificable*, asociado con la forma en la que se encuentran organizadas las unidades espaciales y, especialmente, con el nivel de agregación escogido.

En concreto, para el caso de la I de Moran, *Chou* (1991) demuestra como dicho estadístico se ve influido por los denominados *efectos escala* asociados con cambios tanto en el tamaño del área de estudio como en el nivel de resolución del mapa.

Con relación a este último aspecto, Chou muestra como a medida que incrementa el nivel de desagregación de las unidades espaciales, comienza a dominar un esquema de autocorrelación espacial positiva.

Para poder reforzar más las soluciones obtenidas mediante el contraste la I Ley de Moran, se opta por implementar también este contraste, el del estadístico *G* de Getis y Ord. En este caso, la hipótesis nula también significa ausencia de autocorrelación espacial y el estadístico se calcula como

$$G(d) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(d)x_i x_j}{\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j} i \neq j$$

donde dos pares de regiones *i* y *j* son consideradas vecinas dependiendo de una distancia *d* determinada, que puede estar definida según diferentes criterios como se ha comentado con anterioridad.

Entonces, si Z(G), el contraste de G de Getis y Ord, definido como

$$Z(G) = \frac{G - E(G)}{\sqrt{V(G)}}$$

no es significativo se tiene una distribución espacial aleatoria de la variable, si Z(G) es positivo y significativo significa que hay una concentración de valores elevados en regiones vecinas, pero si Z(G) es significativo y negativo entonces significa que estadísticamente hay una concentración de valores bajos en regiones vecinas.

Hay que destacar que solamente es aplicable a variables positivas (no aplicables a los residuos de una regresión) y con una matriz de pesos simétricas.

Como en el caso anterior, el cálculo de E(G) y E(G^2) se basa en la hipótesis nula de aleatoriedad espacial:

$$E_A[G] = \frac{S_0}{N(N-1)}$$

$$E_A[G^2] = \frac{1}{(m_1^2 - m_2)^2 N^{(4)}}\left[B_0 m_2^2 + B_1 m_4 + B_2 m_1^2 m_2 + B_3 m_1 m_3 + B_4 m_1^4\right]$$

donde,

$$m_j = \sum_{i=1} x_i^j, \quad N^{(r)} = N(N-1)(N-2)...(N-r+1)$$

$$B_o = (N^2 - 3N + 3)S_1 - NS_2 + 3K^2$$

$$B_1 = -\left[(N^2 - N)S_1 - 2NS_2 + 3K^2\right]$$

$$B_2 = -\left[2NS_1 - (N+3)S_2 + 6K^2\right]$$

$$B_3 = 4(N-1)S_1 - 2(N+1)S_2 + 8K^2$$

$$B_4 = S_1 - S_2 + K^2$$

$$K = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$$

$$S_1 = \frac{1}{2}\sum_i\sum_j (w_{ij} + w_{ji})^2 \qquad S_2 = \sum_i\sum_j (w_{i.} + w_{.i})^2 \qquad w_{i.} = \sum_{j=1}^{N} w_{ij}$$

Ante una muestra bastante grande, el contraste de Getis y Ord, Z(G) también sigue una distribución N(0,1).

En este caso los resultados obtenidos confirman que estadísticamente se puede suponer que hay una autocorrelación espacial:

Matriz tipo queen				
Tipos de Pesos	**Estadístico G**	**E(G)**	**V(G)**	**p-valor**
Sin estandarizar	3.579811e-03	2.946817e-03	1.406614e-09	2.2e-16
Estandarizados por filas	5.240493e-04	4.657662e-04	1.991130e-11	2.2e-16

- Moran Scatterplot

Con la gráfica scatterplot se muestra si existe asociación lineal entre la variable de una cierta región *i* y las variables de sus regiones vecinas. Este gráfico, suele ser un gráfico de mucha ayuda para ver la autocorrelación espacial a nivel global y local, pero en este caso al disponer de un número tan alto de observaciones y tener valores tan extremos no es muy buena, por ello no ha sido significativo el resultado obtenido y se ha decidido no incluirla en el análisis por la poca aportación que ofrecía en el estudio.

3.4 Autocorrelación espacial a nivel local

La I ley de Moran (al igual que la G de Getis y Ord) son contrastes globales de autocorrelación espacial. Es decir, suministran un único valor del contraste para toda la muestra y sirven para concluir, en términos medios, cuál es el patrón de comportamiento general de la variable. Pero pueden ocurrir situaciones como las siguientes:

- Distribución aleatoria en el espacio de una variable en términos globales pero existencia de pequeños clusters de regiones con concentraciones de valores elevados (o bajos) de la variable objeto de estudio.
- Esquema de dependencia espacial positiva a nivel global pero existencia de un pequeño esquema centro-periferia alrededor de una región determinada.

Para solucionar esto, se hace uso de los contrastes de autocorrelación espacial a nivel local. Dichos contrastes locales de autocorrelación espacial empleados permiten detectar la presencia de:

- Clusters espaciales: situación en la que una región i y sus regiones vecinas concentran valores especialmente altos o bajos de una variable en comparación con el valor medio esperado.
- Outliers espaciales: situación en la que región i muestra un valor de la variable muy diferente al de sus regiones vecinas (disimilitud significativa entre el valor de la variable en la región i y sus regiones vecinas).

Se han usado dos tipos de contraste: *Local Moran I_i* (Anselin, 1995) y *New G_i* de Ord y Getis* (1995).

- *Local Moran I_i (Anselin, 1995)*

  La hipótesis nula del contraste Local Moran I_i, es la ausencia de autocorrelación espacial alrededor de la región *i*. Su estadístico es el siguiente:

$$I_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2 / N} \sum_{j \in J_i} w_{ij}(x_j - \bar{x})$$

  donde J_i es el conjunto de regiones vecinas a la región i.

  Así, si $Z(I_i)$, definido como

$$Z(I_i) = \frac{I_i - E(I_i)}{\sqrt{V(I_i)}}$$

  es no significativo, significa que hay una distribución aleatoria alrededor de la región *i*. Si $Z(I_i)$ es positivo y significativo, quiere decir que hay un cluster espacial en la región *i* (concentración significativa de valores similares en la región i y en sus vecinas). Si Z(Ii) es significativa y negativa, entonces se tiene un outlier espacial en la región *i* (región *i* muestra valores significativamente diferentes a los de sus vecinas).

  Cabe decir que ante una muestra suficientemente grande, el contraste de la I_i de Moran estandarizado, $Z(I_i)$, sigue una distribución N(0,1).

  Basándose en la hipótesis nula de aleatoriedad se obtiene que $E(I_i)$ y $E(I_i^2)$ son de la forma:

$$E\ (I_i) = -\frac{W_i}{N-1}$$

$$E(I_i)^2 = \frac{W_{i(2)}(N - b_2)}{(N-1)} + \frac{2W_{i(kh)}(2b_2 - N)}{(N-1)(N-2)}$$

Donde W_i es la suma de todos los elementos de **W** de la fila correspondiente a la región *i*, y el resto de variables son:

$$b_2 = \frac{m_4}{m_2^2}, \quad m_t = \frac{\sum_i z_i^t}{N}, \quad W_{i(2)} = \sum_{j \neq i} w_{ij}^2 \;\; y \;\; 2W_{i(kh)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih}$$

- *New G_i* de Ord y Getis (1995)*

En este caso, la hipótesis nula es la misma que la hipótesis nula del contraste local de Moran I_i, pero ahora se tiene el siguiente estadístico:

$$New - G_i^* = \frac{\sum_{j=1}^{N} w_{ij}(d) x_j - W_i^* \bar{x}}{sd(x)\left\{ NS_{1i}^* - W_i^{*2} \Big/ (N-1) \right\}^{1/2}} \quad \forall j$$

Donde,

$$sd(x) = \sqrt{\frac{\sum_j (x_j - \bar{x})^2}{N}}, \;\; S_{1i}^* = \sum_j w_{ij}^2(d), \quad W_i^* = \sum w_{ij}(d)$$

La expresión ha sido estandarizada y se distribuye asintóticamente como una N(0,1).

Si se rechaza la hipótesis nula, dependiendo del estadístico, puede indicar que hay clusters espaciales de valores elevados alrededor de la región *i* (*New-G_i** > 0) o clusters espaciales de valores bajos alrededor de la región *i* (*New-G_i** < 0). Es decir, este contraste aporta una información complementaria, al contraste local de Moran I_i.

Los resultados obtenidos según el *contraste local de Moran Ii* muestran como en general los clusters que se observan son debido a zonas censales vecinas entre sí que tienen valores similares de la SMR, las podemos visualizar en zonas como en el sur de Barcelona, la zona de Sant Adrià del Besòs, la zona nord-oeste de El Prat de Llobregat, Sant Joan Despí, el sur de Sant Feliu de Llobregat (*Figura9*).

Para ver si estos clusters tienen valores altos, o por el contrario son clusters formados por un conjunto de zonas censales que tienen las SMR bajas, se usa el *contraste local de New G_i* de Ord y Getis,* el cual detecta que la mayoría de zonas forman pequeños clusters que contienen una SMR muy alta (*Figura10*).

Figura9
Mapas que representan el valor del estadístico local de Moran para el análisis de la autocorrelación espacial de la SMR

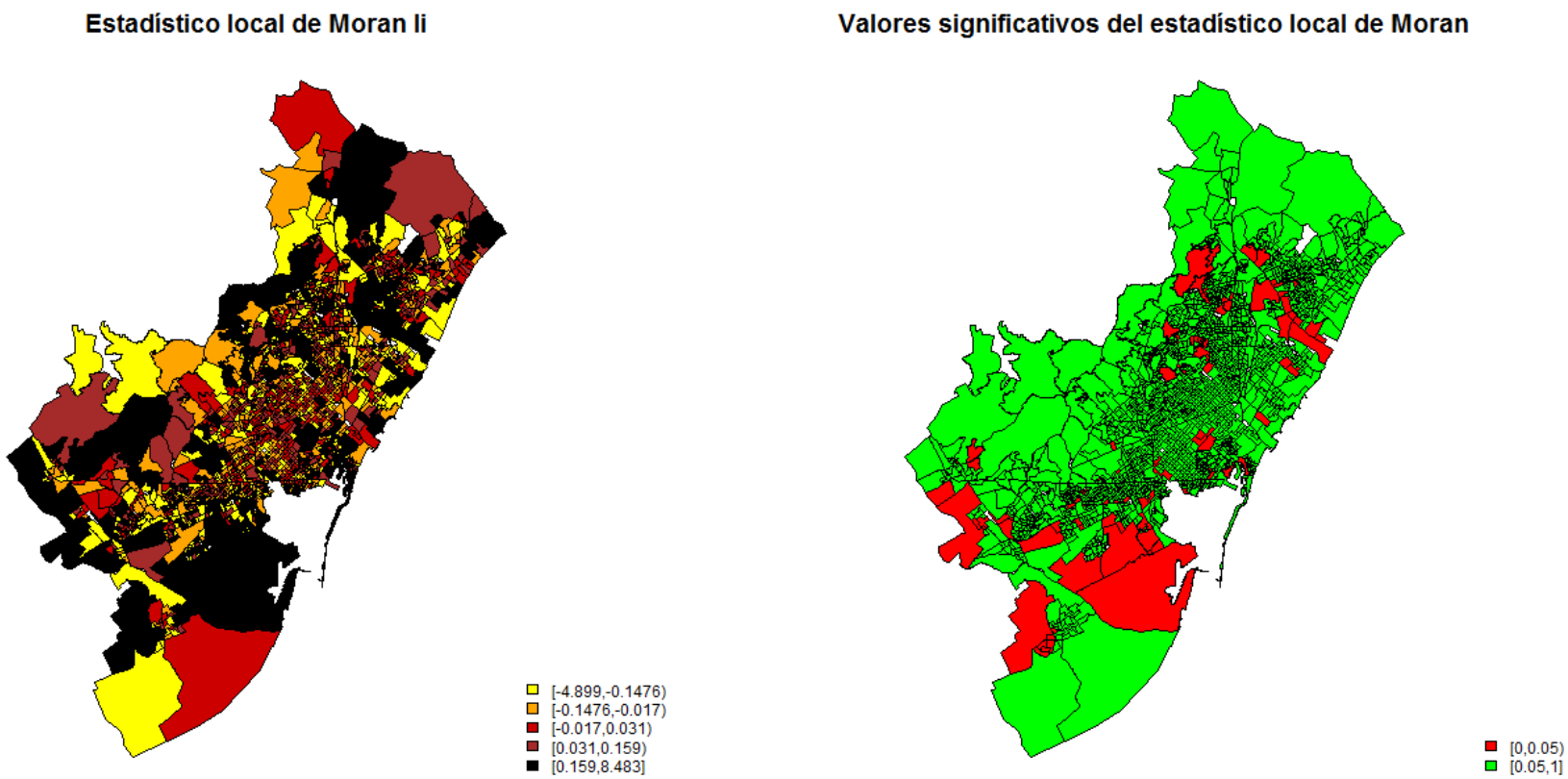
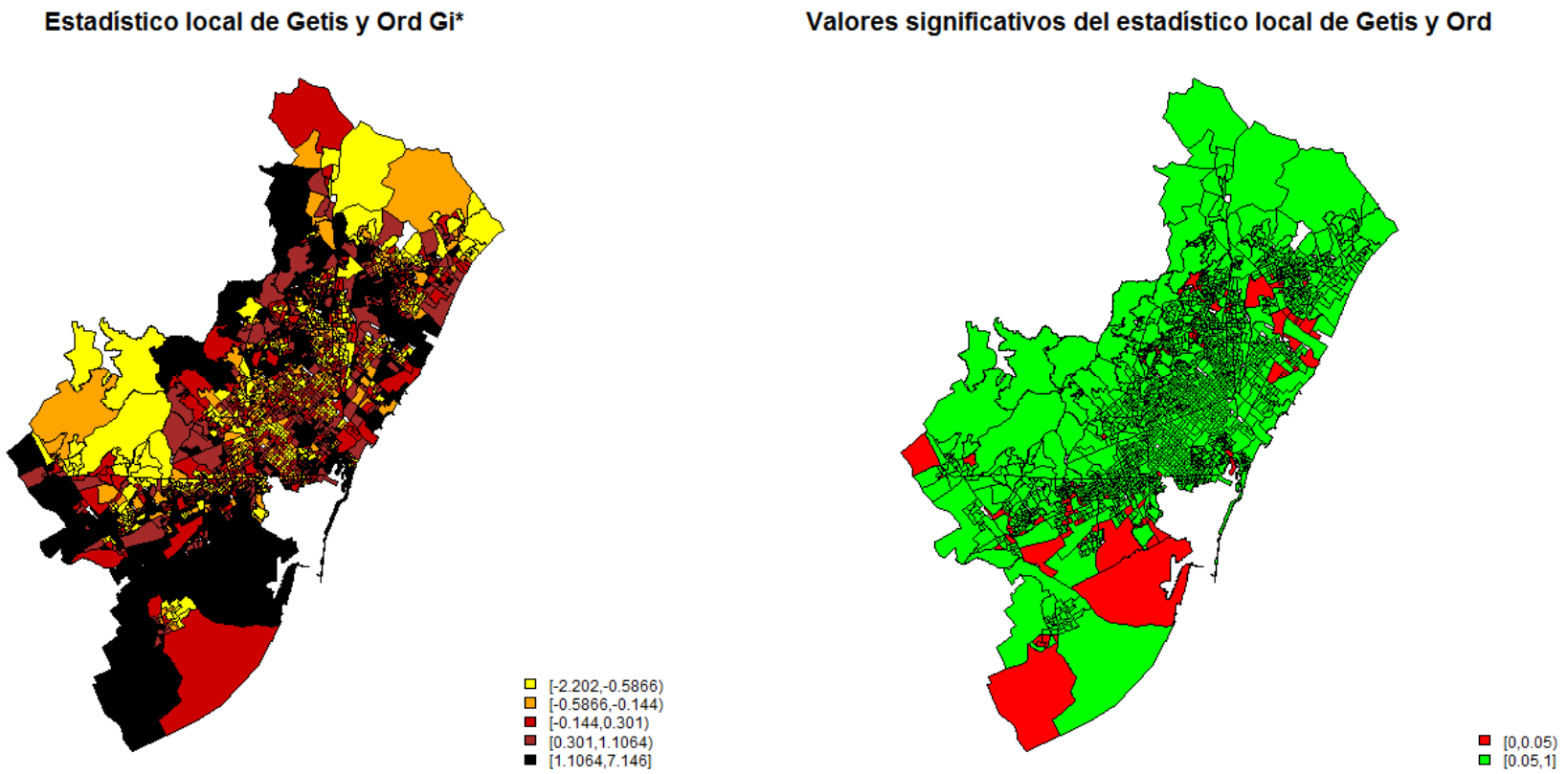


Figura10

Mapas que representan el valor del estadístico local de Getis y Ord para el análisis de autocorrelación espacial de la SMR



3.5 Conclusiones del análisis exploratorio espacial

Una vez analizada a nivel espacial la variable SMR se concluye que estadísticamente existe autocorrelación espacial positiva a nivel gobal entre las diferentes regiones fronterizas, es decir, que regiones próximas tienen valores similares de la tasa de mortalidad estandarizada.
A nivel local se han detectado pequeños clusters en zonas como el sur de Barcelona, Sant Adrià del Besòs, etc. Estos clusters contienen valores altos, es significa que hay una agrupación de regiones con una tasa de mortalidad muy por encima a la media del resto de regiones. Una posible explicación es que las áreas tengan una población pequeña y por tanto su SMR se haya estimado con poca precisión. En esos casos pueden obtenerse valores extremos de SMR por azar. Para prevenir ese tipo de situaciones, en la siguiente sección se realizará un suavizado espacial de los datos.

4. Modelo geoestadístico

4.1 Suavización de la SMR

La SMR depende mucho de la dimensión de la población; su varianza es inversamente proporcional a los valores esperados y eso hace que con una pequeña población obtengamos estimaciones que varían en gran medida. Además, las estimaciones crudas de la SMR en áreas pequeñas son muy inestables y tienen mucho error, es por ello que se opta por suavizar dicha variable, de esta forma se obtienen estimaciones más estables y permite, normalmente, identificar los patrones espaciales de comportamiento de dicha variable en el mapa.

Cuando se habla de suavizar dicha variable, consiste en, cuando hacemos una estimación para un área, utilizar no sólo datos de esa área sino de otras áreas, para tener estimaciones más robustas. Es decir, mediante los valores de la variable en las áreas vecinas, según un suavizado espacial, o las SMR de todas las áreas (la media global), según un suavizado no espacial, se realiza una estimación de la tasa de mortalidad de forma que se obtenga una mayor estabilidad.

La variabilidad en la SMR puede dividirse en dos fuentes:

- *Variabilidad espacial o error estructurado*: esta fuente de variabilidad explica porque las áreas vecinas tienden a tener valores similares. Como se ha observado en el capítulo anterior (*sección 3.5*), existe autocorrelación espacial en nuestros datos. Parte de esta dependencia no es realmente una dependencia estructural, sino que principalmente es debida a variables explicativas que no son incluidas en el análisis y muestran una estructura espacial.

- La segunda fuente es la *variabilidad no estructurada que explica la heterogeneidad entre áreas*. Parte de esta variabilidad puede ser debida a variables sin estructura espacial que no han sido observadas y pueden influir al riesgo relativo.

Existen varios métodos para suavizar la SMR, pero en este trabajo se usará un modelo jerárquico Bayesiano, en el cual se asignan distribuciones a priori a todos los parámetros del modelo. La idea es hacer dos modelos uno de heterogeneidad y otro mixto. El modelo de heterogeneidad no incluye los efectos aleatorios espaciales, en cambio el mixto sí, de esta forma se observará y se comparará el considerar o no considerar dichos efectos.

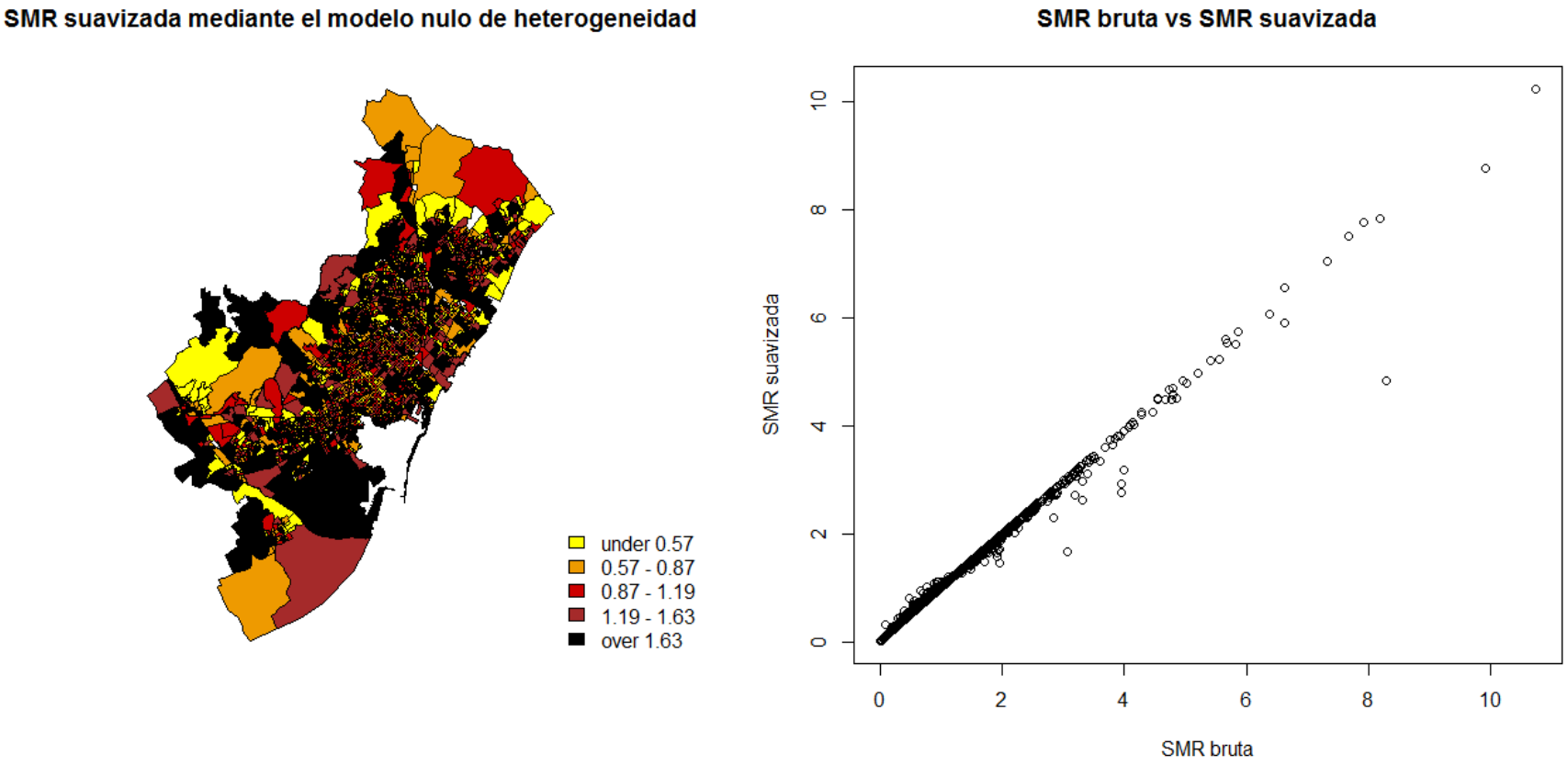
4.2 Modelo de heterogeneidad

El modelo de heterogeneidad con variables Poisson se escribe de la siguiente manera:

$$O_i \sim Poisson(\mu_i \cdot E_i)$$
$$Log(\mu_i) = \beta_0 + U_i$$

Donde O_i denota los casos observados en la zona censal *i*; E_i son los casos esperados en esta zona censal (ver *sección 2.3*); μ_i es la SMR en *i*; U_i los efectos aleatorios no espaciales; y β_0, la constante, que puede ser interpretada como el logaritmo de la SMR basal, es decir la SMR para toda la población. Se asume que la distribución de los efectos aleatorios espaciales sigue una normal de media cero y varianza constante ($U_i \sim N(0,1/\tau_U)$).

Figura11
Mapa que representa la SMR suavizada mediante el modelo de heterogeneidad nulo y gráfico que compara la SMR bruta y suavizada



El modelo de heterogeneidad se completa añadiendo variables explicativas (x_{ij} donde i hace referencia a la sección censal y j define qué variable explicativa representa) y sus correspondientes coeficientes β_{ij}.

$$O_i \sim Poisson(\mu_i \cdot E_i)$$
$$Log(\mu_i) = \beta_0 + \sum_{j=1}^{P} \beta_{ij} \cdot x_{ij} + U_i$$

Donde P, es el número de variables.

En el mapa de la *Figura11*, se puede observar la distribución de la SMR suavizada calculada mediante el modelo sin variables adicionales (modelo nulo, obsérvese la *Tabla6*), en dicho mapa no se identifican muchas diferencias a las SMR brutas. Para ver que sí que existen, en la misma *Figura11*, se comparan las SMR brutas y la SMR estimadas, el que sean casi iguales, es debido a que la mayoría de números de muertos en una zona censal es grande y por lo tanto las estimaciones son estables. En la *Figura12* se puede observar como en las áreas con un número de individuos pequeño se nota la suavización de la estimación, hay que tener en cuenta que los datos de mortalidad total eran datos agregados durante 8 años. De cara a un futuro se plantea estratificar el estudio dependiendo del sexo de las personas, de las causas específicas de mortalidad, etc. Eso hará que se reduzcan mucho los números en cada área y se espera que la suavización mediante el modelo aporte más beneficios.

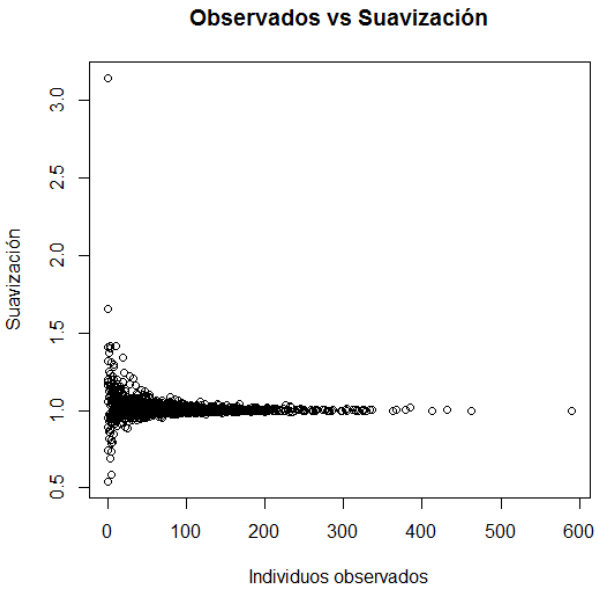


Figura12. Número de individuos observados en cada zona censal vs suavización realizada

4.3 Modelo de Besag, York y Mollié

El modelo mixto usado fue propuesto por Besag, York y Mollié (BYM) en el 1991. Dicho modelo, utiliza la idea del modelo de heterogeneidad anterior, es

decir, que existen unos efectos aleatorios no espaciales que siguen un distribución normal de media cero y varianza constante. Pero, además se le incluyen los efectos de la dependencia espacial, para dichos efectos se utiliza el *modelo condicional autorregresivo (CAR*). Es conocido también como el modelo CAR no intrínseco. Dicho modelo es el modelo más usual y computacionalmente más simple que enfoca la dependencia espacial utilizando promedios de los efectos en zonas vecinas. Por lo tanto el modelo mixto utilizado es:

$$\mathrm{O_i} \sim Poisson(\mu_i \cdot \mathrm{E_i})$$
$$Log(\mu_i) = \beta_0 + \mathrm{U}_i + S_i$$

Donde $\mathrm{O_i}$ denota los casos observados en la zona censal *i*; $\mathrm{E_i}$ son los casos esperados en esta zona censal; μ_i es el riesgo relativo en i; U_i los efectos aleatorios no espaciales los cuales siguen una distribución $\mathrm{U}_i \sim N(0, 1/\tau_{\mathrm{U_i}})$; S_i los efectos aleatorios espaciales; y β_0, la constante, que puede ser interpretado como el logaritmo de la SMR basal.

En el análisis que se realiza en este proyecto los S_i son especificados usando un modelo condicional autorregresivo espacial, el modelo CAR, el cual fue introducido por Besag en 1973, y se utilizan mucho en los contextos de Gibbs Sampling y más generalmente en los métodos de Monte Carlo Markov Chain para el ajuste de modelos jerárquicos. El modelo CAR se define de la siguiente manera:

$$\{S_i = s_i |\, S_j = s_j, j \neq i, j \; es \; un \; vecino \; de \; i, \tau_S\} \sim N(\frac{1}{n_i}\sum_{i \sim j} s_j \; , \frac{1}{n_i \tau_S})$$

El término de los efectos aleatorios asociado con la zona censal *i,* dados todos los términos de los efectos aleatorios de todos los vecinos de *i,* sigue una distribución normal.

El término n_i es el número de vecinos de la zona censal, la relación *i ~ j* indica que las zonas censales *i* y *j* son vecinas. El hiperparámetro de precisión τ_S (se le llama hiperparámetro por ser un parámetro de una distribución a priori) es equivalente a $\sigma_S^2 = 1/\tau_S$, donde σ_S^2 hace referencia a la varianza de los efectos espaciales aleatorios S_i, este parámetro τ_S cumple que $log\tau_S = \varphi$, donde $\varphi \sim loggamma\,(a, b)$.

El modelo BYM se completa añadiendo variables explicativas ($x_{ij,}$ donde *i* hace referencia a la sección censal y *j* define qué variable explicativa representa*) y* sus correspondientes coeficientes β_{ij}:

$$O_i \sim Poisson(\mu_i \cdot E_i)$$

$$Log(\mu_i) = \beta_0 + \sum_{j=1}^{N} \beta_{ij} \cdot x_{ij} + U_i + S_i$$

En la *Figura13*, se puede observar la distribución de la SMR suavizada calculada mediante el modelo sin variables adicionales (modelo nulo, obsérvese *Tabla 7*), donde no se observan muchas diferencias con la SMR bruta, para ver que sí que existen, en la misma *Figura13*, se pueden ver las diferencias entre la SMR bruta y la SMR estimada por este modelo. En la *Figura14* se muestra como, al igual que con el modelo de heterogeneidad, en las áreas con un número de individuos pequeño la suavización de la estimación de la tasa SMR es más notable que en áreas con muchos individuos.

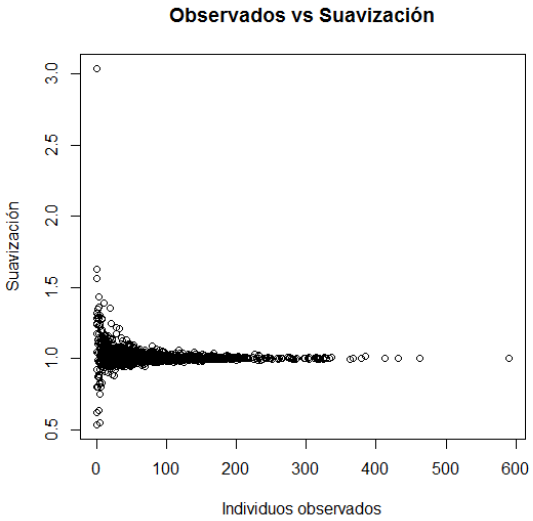


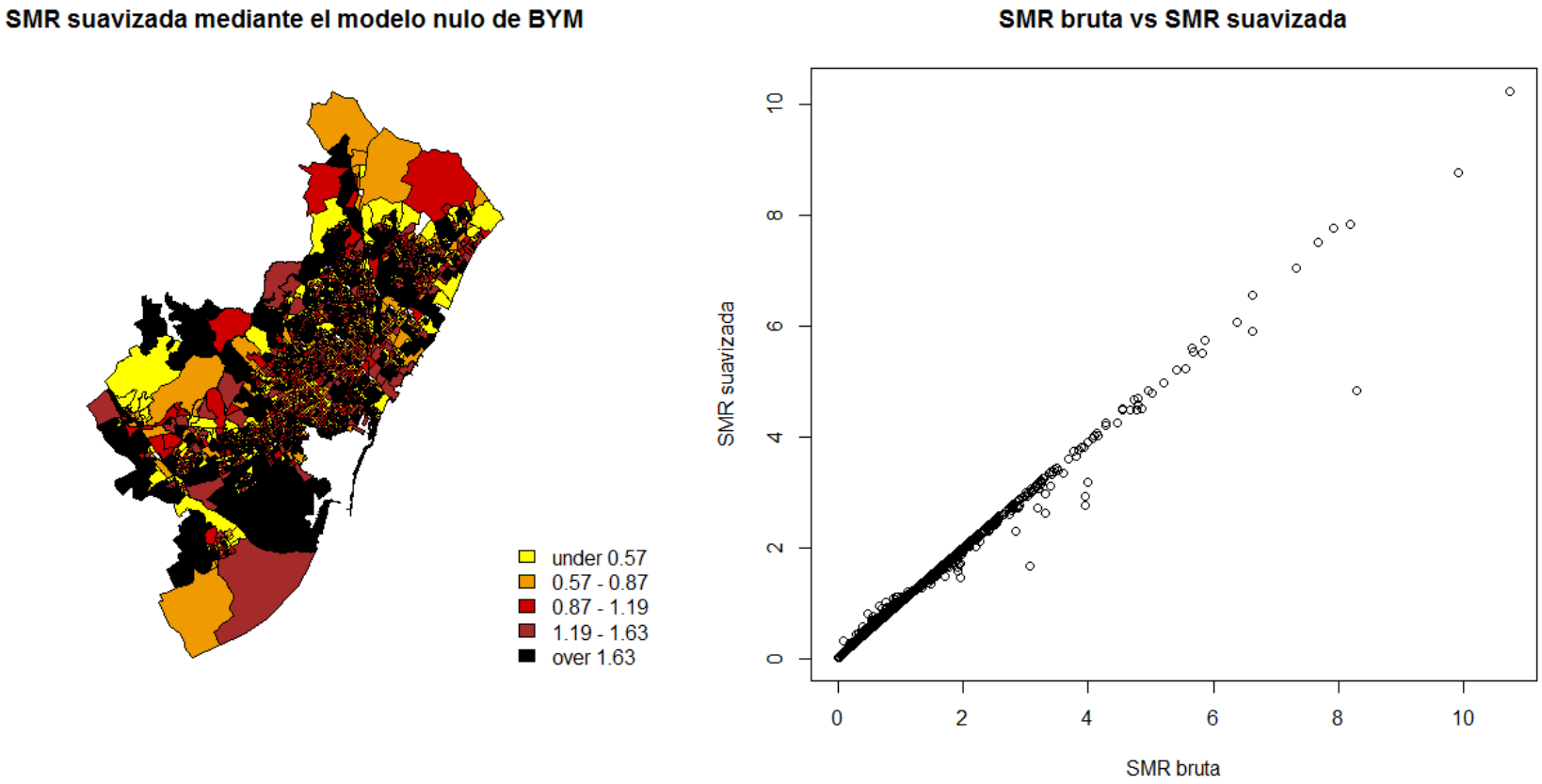
Figura14. Número de individuos observados en cada zona censal vs suavización realizada

4.4 Estimación de un modelo

Una estimación bayesiana requiere que la verosimilitud sea dada, es decir, que el modelo quede especificado y que las distribuciones a priori de los parámetros de interés también.

Una vez el modelo (la verosimilitud) y las distribuciones a priori son especificados, se obtiene la distribución a posteriori. En general, si θ es nuestro parámetro de interés, $f(\theta)$ su distribución a priori y $f(y|\theta)$ la función de verosimilitud de los datos, la distribución a posteriori del parámetro dados

Figura13
Mapa que representa la SMR suavizada mediante el modelo de BYM nulo y gráfico que compara la SMR bruta y suavizada



los datos, que es la utilizada para hacer inferencia en el contexto bayesiano, se puede obtener como $f(\theta|y) \propto f(y|\theta)f(\theta)$. En contextos de múltiples parámetros de interés interesa derivar las distribuciones marginales a posteriori de los parámetros, que requieren de integraciones de la función a posteriori que la mayoría de veces no tienen fórmula cerrada. Para aproximarlas, se pueden utilizar métodos analíticos o métodos de simulación. Los métodos de simulación iterativos conocidos como Markov Chain Monte Carlo (MCMM) son muy utilizados.

Los métodos de simulación iterativos como los implementados en los programas Winbugs o Openbugs requieren un elevado coste computacional (hay que pensar que se necesitan unas 50000 iteraciones, para obtener los parámetros de un modelo donde solamente se incluye una variable) y requieren un control de la convergencia de las simulaciones. Por ese motivo, se decidió optar por un método de estimación diferente, que aproxima las distribuciones marginales a posteriori utilizando métodos de Laplace para las integrales. Así, las integrales son aproximadas por métodos numéricos (sumas finitas). El método de estimación concretamente es llamado *Integrated Nested Laplace Approximations* y está descrito en el siguiente artículo:

2009: Rue H., Martino S. and Chopin N.: Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). Journal of the Royal Statistical Society, Series B, 71, 319–392

Las especificaciones utiizadas en la modelización con INLA fueron:

- $\beta_{ij} \sim N(0,0.0001)$, en este caso no se considera el término independiente.
- S_i, los efectos aleatorios espaciales, siguen una distribución condicionada

  $$\{S_i = s_i | S_j = s_j, j \neq i, j \; es \; un \; vecino \; de \; i, \tau_S\} \sim N(\frac{1}{n_i}\sum_{i\sim j} s_j \; , \frac{1}{n_i \tau_S})$$

  donde el hiperparámetro τ_S se obtiene haciendo una transformación logarítmica, $\varphi = \log(\tau_S)$, donde $\varphi \sim loggama\,(1, 0.00005)$.

- U_i, los efectos aleatorios no espaciales, los cuales siguen una distribución $U_i \sim N(0, 1/\tau_U)$. Donde el hiperparámetro τ_U se define igual que el hiperparámetro τ_S.
- El término β_0, sigue una distribución invariante, en inglés el término se conoce como *"flat prior".* Es decir, que su distribución se considera constante, dando la misma probabilidad a todos los valores.

Las aproximaciones son muy buenas para modelos como los utilizados aquí, donde los efectos aleatorios siguen una distribución Normal. Este método fue implementado en el software R versión 2.15.1 y usando el paquete INLA, el cual nos permite obtener buenas estimaciones en un tiempo muy óptimo.

4.5 Comparación de modelos

Para decidir que variables entran en el modelo final para cada uno de los modelos (el modelo de heterogeneidad y el modelo BYM) se ha decidido seguir los siguientes pasos:

- Primero, crear un modelo univariante añadiendo una sola variable al modelo nulo, con cada una de las variables.
- Segundo, ver que variables son estadísticamente significativas, es decir aquellas que cuando mido en términos de riesgo relativo, su intervalo de credibilidad al 95% no contiene el valor 1.
- Tercero, de los modelos que me han quedado escojo aquel que tenga el DIC (*Deviance Information Criterion*) más bajo.
- Cuarto paso, vuelve a repetir todo el proceso, pero en vez del modelo nulo ahora considero el modelo univariante, y así sucesivamente.

La razón por la cual se escogen a parte de modelos que tengan todas las variables significativas, también que tengan un DIC bajo, es porque el DIC es un valor que me define cómo de explicativo es el modelo, es una especie de generalización jerárquica del AIC (Aikake Information Criterion) y BIC (Bayesian Information Criterion). El DIC es muy utilizado en modelos bayesianos. Para definir el DIC antes hay que estudiar la distribución posterior de la devianza clásica, la cual se define como:

$$D(S, U, \beta) = -2 \cdot ln(f(Y|S, U, \beta)) + 2 \cdot ln(g(Y))$$

Aquí, $f(Y|S,\mathrm{U},\beta)$ es la función de verosimilitud, es decir, la función de densidad condicionada a las observaciones dados los parámetros que se estiman, y $ln(g(Y))$ denota un término totalmente estandarizado (*Y* hace referencia a la variable respuesta del modelo). El DIC, consiste en dos componentes, un término que mide la bondad de ajuste y penaliza los términos que hacen incrementar la complejidad del modelo:

$$DIC = \widehat{D} + p_D$$

El primer término, es una medida Bayesiana de la bondad del modelo, se define como la devianza esperada posterior:

$$\widehat{D} = E_{S,\mathrm{U},\beta|Y}[-2 \cdot \ln\big(f(Y|S,\mathrm{U},\beta)\big)]$$

El modelo que mejor se ajusta a los datos, es aquel que tiene un valor mayor de verosimilitud. Así que si $\widehat{D}$, el cual se define como -2 veces la verosimilitud, tiene valores bajos, entonces el modelo se ajusta mejor.
La segunda componente mida la complejidad del modelo por el *número efectivo de parámetro*, $p_{D,}$ definido como la diferencia entre la esperanza posterior de la devianza y la devianza evaluada en la esperanza posterior $\hat{\theta}$ de los parámetros:

$$p_D = \widehat{D} - D\big(\hat{\theta}\big) = E_{S,\mathrm{U},\beta|Y}[D(\theta)] - D\big(E_{S,\mathrm{U},\beta|Y}[\theta]\big) =$$
$$= E_{S,\mathrm{U},\beta|Y}[-2 \cdot ln(f(Y|S,\mathrm{U},\beta))] + 2 \cdot ln\big(f\big(Y\big|\hat{S},\hat{\mathrm{U}},\hat{\beta}\big)\big)$$

Por definición $-2 \cdot \ln\big(f(Y|S,\mathrm{U},\beta)\big)$ es la información residual en los datos Y condicionados por S,U,β, y se interpreta como una medida de incertidumbre. Además, la última ecuación muestra como p_D puede ser interpretado como el exceso esperado de acierto sobre la estimación residual en los datos Y condicionado por S,U,β. Esto significa que se puede interpretar p_D como la reducción esperada de la incertidumbre debida a la estimación.

4.6 Resultados obtenidos para cada modelo

En el siguiente apartado se mostrarán los resultados obtenidos para cada modelo, para ello se observará su valor DIC y el valor de sus coeficientes expresados en términos de riesgo relativos. Dos ejemplos los vemos con el modelo de heterogeneidad nulo y el de BYM nulo.

Tabla6. Modelo nulo de heterogeneidad

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8994 (0.8686 – 0.9312)
Varianza $\hat{U}$	0.6492 (0.6085 – 0.6898)
DIC	16606.89

Tabla7. Modelo nulo de BYM

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8996 (0.8702 – 0.9299)
Varianza $\hat{U}$	0.5964 (0.5482 – 0.6590)
Varianza $\hat{S}$	0.1240 (0.0514 – 0.3130)
%Variabilidad de los efectos espaciales aleatorios: 17.21%	
DIC	16604.71

En este caso ambos coeficientes basales son significativos ya que el intervalo de credibilidad al 95% no contiene el valor 1. Pero, si se comparan los DIC, el modelo de BYM contiene un valor más pequeño que el modelo nulo de heterogeneidad. Se pueden ver, como al introducir los efectos aleatorios espaciales, los no espaciales reducen su varianza.

- *Modelo de heterogeneidad*

  En la *Tabla 8* se muestra el efecto de cada variable en el modelo univariante de heterogeneidad. Se puede observar, como las variables que miden la concentración de partículas del aire, la variable que mide el nivel de educación o la que indica el porcentaje de hogares en una zona censal, no son variables significativas, es decir no incluyen el uno, por la razón que se comentó en el apartado anterior. La primera variable que se tiene en

cuenta en el modelo de heterogeneidad, en este caso, es el porcentaje edificaciones vacías del área censal ya que tiene un DIC del 16604.87, el más bajo, es decir, de todos los modelos univariantes creados este es el que más se ajusta.

Tabla 8. Modelos univariantes de heterogeneidad

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8655 (0.8284 - 0.9043)
$exp(\hat{\beta}_{Mediana\ NDVI})$[a]	**1.0512 (1.0153 – 1.0885)**
DIC	16606.42
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.0454 (0.9473 - 1.1536)
$exp(\hat{\beta}_{Manuales})$	**0.9969 (0.9950 - 0.9988)**
DIC	16606.84
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9957 (0.9032 - 1.0977)
$exp(\hat{\beta}_{Parados})$	**0.9923 (0.9855 - 0.9992)**
DIC	16606.92
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9411 (0.8822- 1.004)
$exp(\hat{\beta}_{Educación})$	0.9960 (0.9913 – 1.001)
DIC	16607.25
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8381 (0.7985 - 0.8796)
$exp(\hat{\beta}_{Hogar\ unifam})$	**1.0041 (1.0022 – 1.0061)**
DIC	16606.37
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9196 (0.8571 – 0.9866)
$exp(\hat{\beta}_{Hogares})$	0.9995 (0.9982 – 1.0008)
DIC	16606.95
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9668 (0.9052 – 1.0325)
$exp(\hat{\beta}_{Hogares\&Locales})$	**0.9982 (0.9968 – 0.9996)**
DIC	16606.74

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8710 (0.8387 – 0.9046)
$exp(\hat{\beta}_{Locales})$	**1.0309 (1.0164 – 1.0457)**
DIC	16606.21

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.1973 (1.0892 – 1.3161)
$exp(\hat{\beta}_{Alturamediaedif})$	**0.9415 (0.9242 – 0.9591)**
DIC	16605.40

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7963 (0.7301 – 0.8684)
$exp(\hat{\beta}_{Refrig01})$	**1.0063 (1.0022 – 1.0103)**
DIC	16606.92

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7587 (0.6939 – 0.8296)
$exp(\hat{\beta}_{Calef01})$	**1.0038 (1.0020 – 1.0056)**
DIC	16606.18

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7962 (0.7581 – 0.8362)
$exp(\hat{\beta}_{Edifvacia})$	**1.0103 (1.0073 – 1.0132)**
DIC	16604.87

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8691 (0.8036 – 0.9400)
$exp(\hat{\beta}_{encuestaverde})$	1.0009 (0.9991 – 1.0027)
DIC	16606.89

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8039 (0.7234 – 0.8934)
$exp(\hat{\beta}_{Nitrógeno})^{b}$	**1.0161 (1.0020 – 1.0314)**
DIC	16606.84

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.6361 (0.4157 – 0.9734)
$exp(\hat{\beta}_{PM10})$	1.0090 (0.9980 – 1.0200)
DIC	16606.88

	Mediana (95% intervalo de credibilidad)

$e^{\widehat{\beta_0}}$	0.8467 (0.8024 – 0.8934)
$exp(\hat{\beta}_{constr.antes1950})$	**1.0021 (1.0007 – 1.0035)**
DIC	16606.76
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.1114 (1.0241 – 1.2061)
$exp(\hat{\beta}_{constr.despues.1950})$	**0.9964 (0.9951 – 0.9977)**
DIC	16605.95

[a] El coeficiente $\hat{\beta}_{Mediana\ NDVI}$ tiene como unidades el canvio de pasar del primer cuartil al tercer cuartil.

[b] El coeficiente $\hat{\beta}_{Nitrógeno}$ tiene como unidades 10 µg/m^3.

El modelo de heterogeneidad final ajustado, siguiendo el procedimiento descrito en la *sección 4.5* es:

Tabla 9. Modelo final de heterogeneidad

	Mediana **(95% intervalo de credibilidad**)
$e^{\widehat{\beta_0}}$	**0.8142 (0.7123 - 0.9307)**
$exp(\hat{\beta}_{Edifvacia})$	**1.0079 (1.0048 – 1.0110)**
$exp(\hat{\beta}_{Alturamediaedif})$	**0.9303 (0.9115 – 0.9494)**
$exp(\hat{\beta}_{Nitrógeno})$[a]	**1.0410 (1.0244 – 1.0578)**
$exp(\hat{\beta}_{locales})$	**1.0167 (1.0016 – 1.0321)**
$exp(\hat{\beta}_{Mediana\ NDVI})$	**1.0762 (1.0395 – 1.1142)**
Varianza $\widehat{U}$	0.6142 (0.5757 - 0.6528)
DIC	16601.97

[a] El coeficiente $\hat{\beta}_{Mediana\ NDVI}$ tiene como unidades el canvio de pasar del primer cuartil al tercer cuartil.

[b] El coeficiente $\hat{\beta}_{Nitrógeno}$ tiene como unidades 10 µg/m^3.

Según los resultados obtenidos, por cada unidad de aumento en la media de plantas de los edificios, la mortalidad en la zona censal se multiplica por 0.9304, es decir se reduce. Por tanto, son las áreas con edificios más altos donde la tasa de mortalidad es menor. En el caso del dióxido de nitrógeno por cada incremento de 10 µg/m^3, el riesgo relativo aumenta un 4.1%. Por tanto, en las áreas más contaminadas la mortalidad es mayor. Así, si se aumenta un 1% la cantidad de edificaciones vacías o locales la tasa de mortalidad aumenta un 1.0079 veces o un 1.0167 veces, respectivamente.

En el caso, del indicador NDVI (el cual me permite medir la cantidad de vegetación de un área censal), el incremento de tener un índice que corresponde al primer cuartil a tner un índice que corresponde al tercer cuartil corresponde a un riesgo relativo 1.0762 veces mayor, parece ser entonces que las zonas censales con mucha vegetación están asociadas con tener una tasa de mortalidad más alta, en consecuencia en zonas de poca vegetación la tasa de mortalidad es más baja. Por último, hay que observar que el DIC tampoco ha disminuido considerablemente, esto es debido a que seguramente las variables explican un porcentaje muy pequeño de la variabilidad de la SMR.

- *Modelo BYM*

En los modelos univariantes de BYM (*Tabla 10*) se observa que las variables que se pueden considerar no significativas son las relacionadas con: el nivel de concentración de dióxido de nitrógeno de la zona censal y de concentración de partículas, la variable correspondiente a la encuesta realizada en la zona censal sobre la cantidad de verde que percibe un individuo allá donde viva y la variables que tienen el porcentaje del nivel de educación, el porcentaje de hogares de la zona censal y el porcentaje de construcciones realizadas antes 1950.

Por otra parte, la variable que proporciona en el modelo univariante un DIC más pequeño, DIC = 16600.74, es la relacionada con la altura media de los edificios de la zona censal, es decir está es la primera variable en entrar en el modelo final.

Tabla 10. Modelos BYM univariantes

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8188 (0.7805 - 0.8589)
$exp(\hat{\beta}_{Mediana\ NDVI})^a$	**1.1305 (1.0796 – 1.1847)**
DIC	16602.42
	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.2566 (1.0450 - 1.5236)
$exp(\hat{\beta}_{Manuales})$	**0.9932 (0.9893 - 0.9969)**
DIC	16603.83

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.0308 (0.9254 – 1.1480)
$exp(\hat{\beta}_{Parados})$	**0.9897 (0.9820 – 0.9975)**
DIC	16604.47

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9503 (0.8806 – 1.0258)
$exp(\hat{\beta}_{Educación})$	0.9952 (0.9892 – 1.0012)
DIC	16605.00

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7813 (0.7420 – 0.8228)
$exp(\hat{\beta}_{Hogar\ unifam})$	**1.0083 (1.0059 – 1.0107)**
DIC	16601.99

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8625 (0.7961 – 0.8626)
$exp(\hat{\beta}_{Hogares})$	1.0009 (0.9994 – 1.0025)
DIC	16604.60

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.0788 (1.0002 – 1.1637)
$exp(\hat{\beta}_{Hogares\&Locales})$	**0.9955 (0.9938 – 0.9972)**
DIC	16602.83

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8747 (0.8387 – 0.8430)
$exp(\hat{\beta}_{Locales})$	**1.0270 (1.0116 – 1.0426)**
DIC	16604.52

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.4724 (1.3117 – 1.6541)
$exp(\hat{\beta}_{Alturamediaedif})$	**0.9014 (0.8803 – 0.9227)**
DIC	16600.74

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7327 (0.6532 – 0.8204)
$exp(\hat{\beta}_{Refrig01})$	**1.0106 (1.0050 – 1.0163)**
DIC	16604.00

	Mediana (95% intervalo de credibilidad)

$e^{\widehat{\beta_0}}$	0.6214 (0.5457 – 0.7064)
$exp(\hat{\beta}_{Calef01})$	**1.0082 (1.0055 – 1.0111)**
DIC	16601.54

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.7836 (0.7459 – 0.8232)
$exp(\hat{\beta}_{Edifvacia})$	**1.0116 (1.0085 – 1.0148)**
DIC	16602.26

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9275 (0.8447 – 1.0199)
$exp(\hat{\beta}_{encuestaverde})$	0.9992 (0.9969 – 1.0015)
DIC	16604.68

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.9119 (0.7819 – 1.0668)
$exp(\hat{\beta}_{Nitrógeno})$[b]	0.9980 (0.9763 – 1.0201)
DIC	16604.75

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8787 (0.5371 – 1.4426)
$exp(\hat{\beta}_{PM10})$	1.0006 (0.9879 – 1.0133)
DIC	16604.79

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	0.8695 (0.8141 – 0.9289)
$exp(\hat{\beta}_{constr.antes1950})$	1.0012 (0.9992 – 1.0031)
DIC	16604.85

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	1.1702 (1.0505 – 1.1700)
$exp(\hat{\beta}_{constr.despues.1950})$	**0.9955 (0.9938 – 0.9973)**
DIC	16603.94

[a] El coeficiente $\hat{\beta}_{Mediana\ NDVI}$ tiene como unidades el canvio de pasar del primer cuartil al tercer cuartil.

[b] El coeficiente $\hat{\beta}_{Nitrógeno}$ tiene como unidades 10 µg/m$^{3.}$

El modelo final obtenido es el de la *Tabla 11*, donde se muestran las variables explicativas que incluye el modelo: la altura media de los edificios (donde por cada unidad de aumento en la media de plantas de los edificios,

la mortalidad en la zona censal se multiplica por 0.9134, es decir se reduce), el porcentaje de edificaciones con calefacción en el 2001 (donde por cada 1% que aumenta el número de calefacciones del área censal se aumenta el riesgo relativo en 1.0079 veces), el porcentaje de edificaciones vacías (en este caso, por cada uno 1% que aumenta el número de edificaciones vacías del área censal aumenta el riesgo relativo en 1.0086 veces) y por último el pasar de tener un valor del índice de vegetación correspondiente al primer cuartil a tener uno correspondiente al tercer cuartil tiene un riesgo relativo 1.1237 veces mayor.

Tabla 11. Modelo BYM final

	Mediana (95% intervalo de credibilidad)
$e^{\widehat{\beta_0}}$	**0.8020 (0.6726 - 0.9559)**
$exp(\hat{\beta}_{Alturamediaedif})$	**0.9134 (0.8914 – 0.9357)**
$exp(\hat{\beta}_{Calef01})$	**1.0079 (1.0050 – 1.0107)**
$exp(\hat{\beta}_{Edifvacia})$	**1.0086 (1.0054 – 1.0119)**
$exp(\hat{\beta}_{Mediana\ NDVI})$[a]	**1.1237 (1.0724 – 1.1779)**
Varianza $\widehat{U}$	0.4527 (0.4027 - 0.5125)
Varianza $\hat{S}$	0.4527 (0.3478 – 0.7597)
%Variabilidad de los efectos espaciales aleatorios: 52.91%	
DIC	16594.59

[a] El coeficiente $\hat{\beta}_{Mediana\ NDVI}$ tiene como unidades el canvio de pasar del primer cuartil al tercer cuartil.

En este caso el DIC tampoco ha disminuido mucho, es de un valor del 16594.59.

Para ver el porcentaje de variabilidad explicada por el término espacial se hace el siguiente cálculo:

$$\frac{\hat{V}(\hat{S})}{\hat{V}(\hat{S}) + \hat{V}(\hat{U})}$$

Donde $\hat{V}(\hat{S})$ y $\hat{V}(\hat{U})$ hacen referencia a las estimaciones de las varianzas de las estimaciones de los efectos aleatorios espaciales y efectos aleatorios no espaciales, respectivamente. De manera que en este caso, se ha obtenido

un valor del 52.91%, es decir más de la mitad de la devianza explicada del modelo es debida a los efectos espaciales aleatorios, si lo comparamos con el modelo nulo que solamente tenía un 17.21%, podríamos decir que ahora el porcentaje se ha triplicado. Si comparamos con las variabilidades estimadas en el modelo nulo, la varianza no espacial se ha reducido. Sin embargo, la varianza asociada al término espacial ha aumentado. Es posible que las variables en el modelo expliquen parte de las diferencias observadas en el SMR, y que una vez tenidas en cuenta, los residuos del modelo presenten una estructura espacial más clara que antes. Eso podría explicar por qué la varianza de la componente espacial aumenta.

Aunque en los resultados finales no se observen diferencias significativas (véase *Figura13*), sí que es verdad que el modelo BYM hace una mejor suavización de la SMR que el de heterogeneidad (*Figura 14*).

También hay que tener en cuenta que el modelo BYM presenta un DIC más bajo que el modelo de heterogeneidad, por lo tanto si hay que decantarse por uno de los dos se optaría por el modelo BYM. Además, el número de variables del modelo BYM es más reducido, una menos, y se ha observado con el cálculo de la variabilidad explicada que gran parte de los efectos aleatorios del modelo son debidos a los efectos aleatorios espaciales, por lo tanto hay que tenerlos presentes, seguramente esto sea debido a variables con estructura espacial que no han sido incluidas en el modelo.

Hay que tener en cuenta que las variables que utiliza cada modelo no son todas las mismas, ambos contienen como variables explicativas: el índice de vegetación NDVI, el porcentaje de altura media de los edificios y el porcentaje de edificaciones vacías. Pero, el modelo BYM, contiene la variable explicativa referente al porcentaje de hogares con calefacción en el 2001, en cambio el de heterogeneidad contiene las variables concentración de dióxido de nitrógeno y el porcentaje de locales. Es decir, los dos modelos finales obtenidos son diferentes. Una posible explicación para las diferencias de variables, es que la componente espacial del modelo puede capturar mejor la variabilidad que en el modelo de heterogeneidad. En tal caso es muy posible que la asociación con el dióxido de nitrógeno no fuera real sino debida a una variable no observada que tenga una cierta correlación con el dióxido de nitrógeno.

Figura13
Mapa que representa la SMR suavizada mediante el modelo final de heterogeneidad y el modelo final de BYM

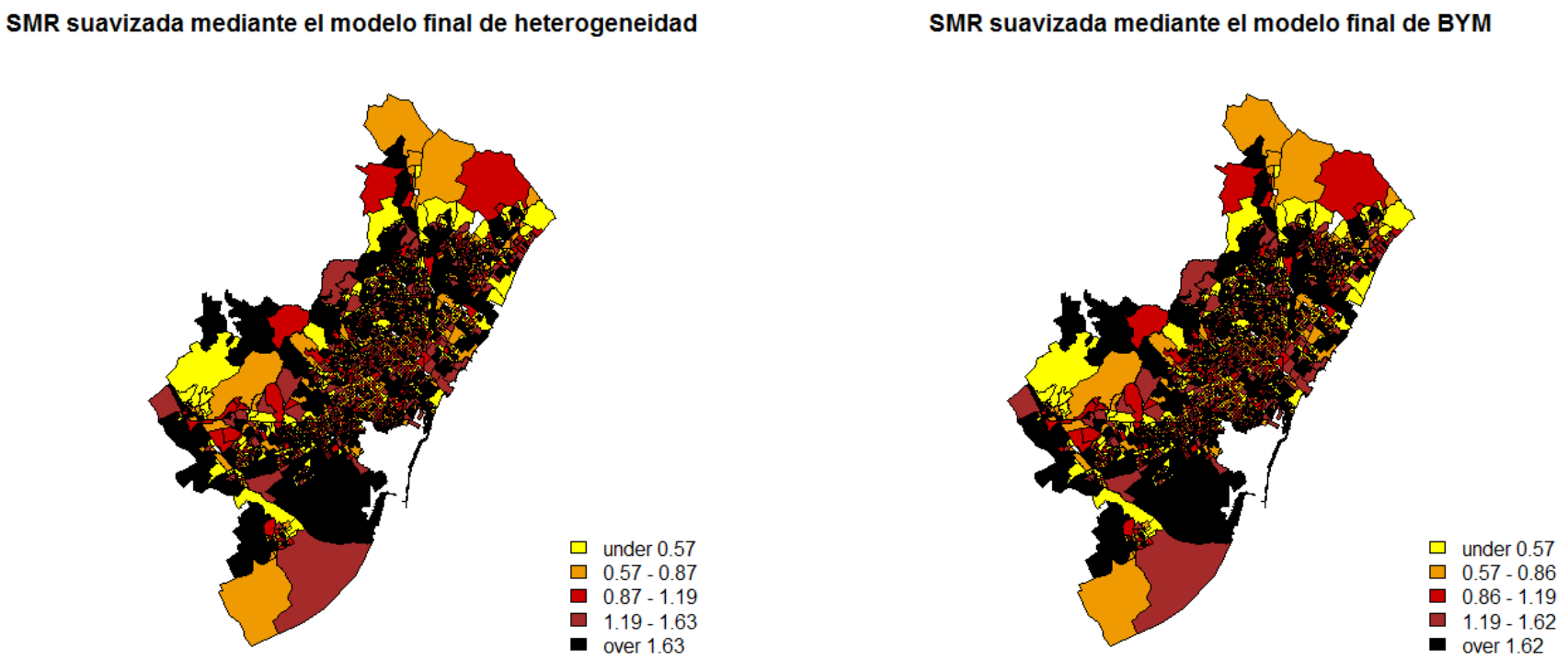
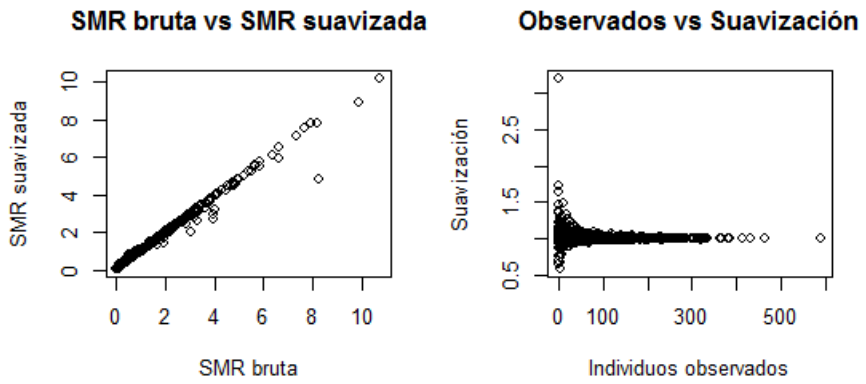
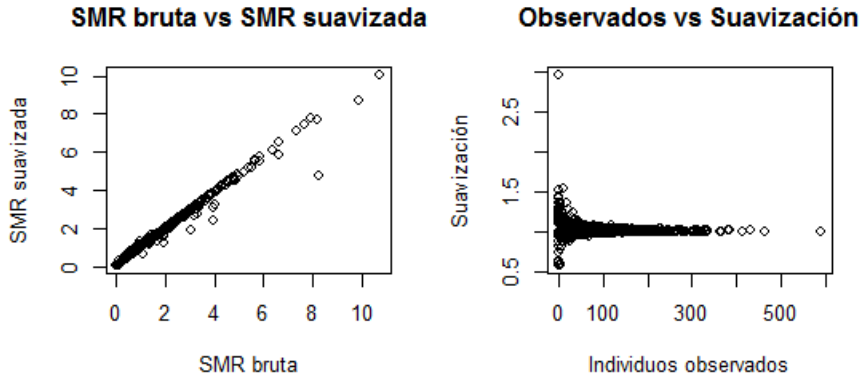


Figura14

Gráfico que compara la SMR suavizada por el modelo final de heterogeneidad y por el modelo final de BYM



MODELO AJUSTADO DE HETEROGNEIDAD



MODELO AJUSTADO DE BYM

En conclusión, dado que relacionada con la aleatoriedad espacial tiene un gran peso, es natural, que los resultados obtenidos con un modelo que tenga en cuenta la espacialidad de los datos sea mucho mejor que no uno que no los tengas, es decir en este caso, habría que decantarse por el modelo BYM.

5. Conclusiones

5.1 Conclusiones y consideraciones finales

Este proyecto se ha basado en un estudio ecológico (los datos no son a nivel individual si no agregado a nivel de sección censal) transversal (basados en un simple punto en el tiempo), dichos estudios presentan, en general, los siguientes inconvenientes:

- Imposibilidad de determinar si la exposición precede a la enfermedad, es decir, imposibilidad para establecer direccionalidad de asociaciones.
- La información de la exposición es muy vulnerable a errores de medición.
- Imposibilidad de identificar relaciones causales entre los factores estudiados, puesto que mide simultáneamente efecto y exposición.

Frente a estas ventajas:

- Permiten estudiar varias variables resultado como enfermedad y exposición.
- Buen control de la selección de los sujetos de estudio.
- Poco tiempo de ejecución del estudio puesto que no hay seguimiento de los individuos y generalmente poco coste económico.

Los resultados de este estudio deben interpretarse con cautela, principalmente por la naturaleza observacional y ecológica del estudio. Por otra parte se han utilizado variables que no tienen porqué ser las mejores para representar el nivel económico, índice de vegetación o de contaminación. Así por ejemplo, como índice de referencia económico hubiera sido mejor un índice de renta per cápita o tener en cuenta otros elementos de contaminación como la cantidad de automóviles que circulan por esa zona, es decir, un medidor de tránsito. El índice de vegetación empleado, NDVI, plantea el inconveniente de ser sensible

a la reflectividad del suelo sobre el que se sitúa la planta. Imaginemos una zona con baja densidad de vegetación. La reflectividad correspondiente a un pixel en la banda infrarroja y en la banda roja, vendría determinado fundamentalmente por el suelo, con una pequeña variación debida a la presencia de vegetación. El resultado es que daría resultados muy similares a los del suelo desnudo y sería imposible detectar la presencia de vegetación. De hecho, este problema es bastante grave cuando la cubierta vegetal es menor del 50%, lo que ocurre bastante a menudo incluso en zonas cultivadas. Para resolver este problema y poder estudiar la vegetación incluso en zonas de baja densidad, se han desarrollado una serie de índices de vegetación que tienen en cuenta la reflectividad del suelo y cuyo objetivo es aislar la información de la vegetación de la que procede del suelo que está bajo ella, estos índices son el PVI (*Perpendicular Vegetation Index*) y el SAVI (*Soil-Adjusted Vegetation Index*). Sería recomendable repetir los análisis con esos índices si pudieran ser obtenidos para las zonas censales del área metropolitana de Barcelona.

Al realizar los modelos, muchos de los coeficientes de las variables asociadas han tenido un valor contrario a lo esperado, cabe destacar que la gran mayoría de variables trabajadas en estudio han sido no significativas en los modelos univariantes, inclusive las variables asociadas al modelo ajustado final eran estadísticamente significativas, pero con valores muy cercanos a uno (en términos de riesgo relativo). Una de las explicaciones se debe a que como se ha comentado con anterioridad las variables no sean las más adecuadas, así por ejemplo, la variable que hace referencia a la altura media de los edificios de la zona censal parece que es una variable que ejerce un valor protector, puesto que a edificios más altos el riesgo relativo es más bajo, cosa que lo que uno espera es lo contrario, ya que si una zona censal tiene un valor medio de altura de edificios muy alto se asocia a una ciudad grande y con un riesgo relativo mayor. En cambio el índice de vegetación mediante la variable NDVI, es un ejemplo de caso contrario, lo esperado hubiese sido que zonas con mucha vegetación tuvieran un riesgo relativo menor que las zonas con poca vegetación, y ha sucedido lo contrario. Es decir, los resultados a la hora de crear el modelo no han sido los esperados, aún haberse confirmado la hipótesis de que sí existe autocorrelación espacial y haberse verificado que un modelo que tiene en cuenta los efectos aleatorios espaciales es mejor que uno que no los tiene en cuenta, los valores de las de los coeficientes de la variables explicativas usadas para crear los modelos de heterogeneidad y BYM no han

sido los valores que uno espera. Una de las posibles causas sea la de estratificación de los datos.

En conclusión, en este proyecto se ha mostrado que existe una autocorrelación espacial entre las diferentes zonas censales del área metropolitana de Barcelona, en lo que a la tasa de mortalidad se refiere. Por otra parte, también se ha visto que a la hora de crear un modelo para estimar las tasas de mortalidad estandarizadas una buena idea sería emplear un modelo que considere los efectos aleatorios como el de BYM. No obstante, los resultados no son del todo buenos y es que de cara al futuro convendría hacer algún tipo de estratificación, es decir hubiera sido mejor trabajar con los datos desagregados debido a sus valores tan extremos y diferentes.

5.2 Consideraciones de cara a un futuro proyecto

De cara a un futuro proyecto, una de las ideas que se proporciona una vez finalizado este estudio es el de la estratificación de los datos, ya sea por sexo o causa de muertes, que son las estratificaciones más comunes, ejemplos de estudios de mortalidad donde se haya realizado una estratificación se puede encontrar en el artículo de Maria Antònia Barceló, Marc Saez y Carme Saurina, *Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain,* en este artículo se realiza un estudio de cómo influye el nivel socioeconómico de las diferentes áreas metropolitanas de Barcelona en la mortalidad de los individuos, y para ello se hace una estratificación de los datos según el sexo de los individuos y según la causa de la muerte: enfermedades respiratorias crónicas (excepto asma); enfermedades del corazón; cáncer de tráquea, bronquios y pulmón; cáncer de vejiga; y linfomas. En los resultados de dicho artículo, se puede observar que en la mayoría de casos como la mortalidad por cáncer de laringe, cáncer de vejiga… los valores de las tasas de mortalidad suavizadas obtenidas muestran que según el sexo se tienen valores diferentes, así el cáncer de laringe, por ejemplo, allá donde los hombres tienen un valor elevado, las mujeres tienen un valor bajo y viceversa.

Otro ejemplo, lo presenta el artículo de Thomas Waldhoer, Martin Wald y Harald Heinzl: *Analysis of the spatial distribution of infant mortality by cause of death in Austria in 1984 to 2006*, en este caso se estudia la mortalidad infantil a nivel espacial en Austria dependiendo de la causa de la muerte: infecciones,

enfermedades respiratorias; problemas perinatales; inmadurez; malformaciones; Síndrome de Muerte Repentina Infantil (*SIDS, Sudden Infant Death Syndrom);* y otras causas*.* En este caso, un ejemplo claro donde se ve la necesidad de la estratificación es en las diferencias entre la mortalidad infantil por SIDS (los valores elevados están en el sudoeste de Austria y el resto son bajos) o por problemas perinatales (los valores elevados están en el nordeste de Austria y el resto son moderados).

Haciendo el estudio de la mortalidad total, se pueden perder patrones, por ejemplo es probable que la asociación de la mortalidad debida al cáncer y la asociación de la mortalidad debida a insuficiencias respiratorias sigan una distribución y un patrón de comportamiento totalmente distinto que induzca a favorecer mejores resultados en un estudio por separado.

Además una opción que mejoraría la veracidad de los resultados es usar otro tipo de indicadores y/o datos, como se ha comentado en la *sección 5.1*.

Hay que tener en cuenta, que se trabajan con muchos datos, por eso al estratificar reducimos el tamaño (los datos totales se reparten entre diferentes estratos), esto quiere decir que los modelos de suavización implementados en este proyecto serán de gran importancia como se pudo ver en la *sección 4.1*, donde se observó como la suavización de la tasa de mortalidad era más significativa en zonas donde el número de individuos fallecidos era reducido. Por eso, en este caso, al estratificar se divide la muestra por grupos, esto quiere decir que el número de individuos fallecidos también, con lo cual mejora la suavización.

APÉNDICE A

Tabla.A1
Número de áreas censales, distribución de la población por área censal y población total de los municipios.

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Badalona	0-4	157	61	37	40	73	9,635
	5-9	157	57	31	37	70	8,950
	10-14	157	63	33	39	76	9,894
	15-19	157	80	42	50	99	12,602
	20-24	157	115	51	78	146	18,060
	25-29	157	124	51	90	149	19,401
	30-34	157	111	59	77	129	17,348
	35-39	157	100	52	68	116	15,680
	40-44	157	92	50	59	111	14,479
	45-49	157	93	53	57	119	14,638
	50-54	157	89	39	60	109	13,921
	54-59	157	77	30	54	98	12,083
	60-64	157	59	23	42	75	9,239
	65-69	157	61	25	43	75	9,514
	70-74	157	53	23	36	65	8,261
	75-79	157	38	18	27	50	6,043
	80-84	157	23	12	14	28	3,534
	85-90	157	11	7	6	14	1,744
	+90	157	5	4	2	7	810
	Total	157	1,311	483	977	1,618	205,836
Barcelona	0-4	1491	40	22	27	47	59,533
	5-9	1491	38	21	25	45	55,929
	10-14	1491	39	23	24	47	57,650
	15-19	1491	47	29	29	56	69,932
	20-24	1491	68	38	46	80	101,330
	25-29	1491	85	40	60	99	126,134

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Barcelona	30-34	1491	79	37	57	92	117,741
	35-39	1491	77	36	55	88	114,330
	40-44	1491	72	39	49	84	106,812
	45-49	1491	65	39	43	76	97,386
	50-54	1491	66	38	43	79	98,498
	54-59	1491	63	32	42	75	93,419
	60-64	1491	53	24	37	64	79,331
	65-69	1491	60	27	43	72	90,050
	70-74	1491	57	25	40	68	84,406
	75-79	1491	47	21	33	58	70,697
	80-84	1491	30	14	20	37	44,046
	85-90	1491	17	9	10	21	24,787
	+90	1491	8	5	4	10	11,873
	Total	1491	1,009	410	746	1,166	1,503,894
Cornellà de Llobregat	0-4	70	52	40	27	47	3,674
	5-9	70	47	27	25	45	3,268
	10-14	70	47	25	24	47	3,295
	15-19	70	58	29	29	56	4,053
	20-24	70	89	34	46	80	6,243
	25-29	70	113	44	60	99	7,934
	30-34	70	104	68	57	92	7,280
	35-39	70	88	51	55	88	6,170
	40-44	70	76	43	49	84	5,338
	45-49	70	67	36	43	76	4,705
	50-54	70	75	35	43	79	5,273
	54-59	70	77	29	42	75	5,362

c

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Cornellà de Llobregat	60-64	70	64	20	37	64	4,473
	65-69	70	64	18	43	72	4,478
	70-74	70	49	20	40	68	3,419
	75-79	70	35	14	33	58	2,471
	80-84	70	21	9	20	37	1,468
	85-90	70	10	5	10	21	721
	+90	70	5	4	4	10	354
	Total	70	1,143	419	865	1,332	79,979
Esplugues de Llobregat	0-4	29	59	19	46	71	1,709
	5-9	29	62	18	48	75	1,802
	10-14	29	71	22	55	82	2,069
	15-19	29	97	37	69	126	2,825
	20-24	29	135	49	108	158	3,903
	25-29	29	143	45	116	167	4,136
	30-34	29	113	39	91	141	3,289
	35-39	29	108	33	81	129	3,127
	40-44	29	113	39	90	136	3,287
	45-49	29	116	41	86	146	3,359
	50-54	29	119	42	89	139	3,456
	54-59	29	111	36	94	130	3,225
	60-64	29	86	31	64	106	2,485
	65-69	29	78	30	63	98	2,265
	70-74	29	56	23	47	71	1,633
	75-79	29	42	15	31	52	1,224
	80-84	29	26	11	19	31	753

d

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Esplugues de Llobregat	85-90	29	14	8	8	18	406
	+90	29	6	3	4	7	174
	Total	29	1,556	430	1,308	1,804	45,127
L'Hospitalet de Llobregat	0-4	226	41	21	29	49	9,271
	5-9	226	37	20	25	45	8,385
	10-14	226	42	24	28	49	9,544
	15-19	226	55	35	34	65	12,403
	20-24	226	87	47	56	104	19,765
	25-29	226	103	41	71	125	23,185
	30-34	226	86	34	63	103	19,415
	35-39	226	75	31	53	91	17,021
	40-44	226	68	35	47	79	15,279
	45-49	226	66	46	38	79	14,958
	50-54	226	73	44	41	91	16,390
	54-59	226	75	37	45	98	16,872
	60-64	226	62	25	43	76	14,034
	65-69	226	64	23	47	76	14,427
	70-74	226	51	19	36	62	11,423
	75-79	226	37	15	27	47	8,436
	80-84	226	21	10	14	27	4,781
	85-90	226	10	6	7	13	2,333
	+90	226	5	3	3	6	1,079
	Total	226	1,058	409	787	1,241	239,019
Montcada i Reixach	0-4	15	96	48	67	105	1,437
	5-9	15	91	46	63	101	1,358
	10-14	15	97	43	72	18	1,461
	15-19	15	116	48	86	129	1,744
	20-24	15	144	57	105	167	2,157

e

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Montcada i Reixach	25-29	15	180	108	111	191	2,705
	30-34	15	170	98	111	164	2,546
	35-39	15	157	75	109	170	2,347
	40-44	15	149	68	101	175	2,229
	45-49	15	123	52	84	142	1,849
	50-54	15	107	42	79	117	1,604
	54-59	15	98	42	67	128	1,467
	60-64	15	80	28	62	103	1,203
	65-69	15	91	29	69	119	1,367
	70-74	15	78	31	59	100	1,172
	75-79	15	57	20	49	67	858
	80-84	15	31	13	24	40	469
	85-90	15	14	7	9	19	215
	+90	15	7	3	6	9	107
	Total	15	1,886	683	1,396	2,048	28,295
El Prat de Llobregat	0-4	36	82	30	62	95	2,946
	5-9	36	75	30	55	92	2,709
	10-14	36	87	32	59	108	3,141
	15-19	36	107	41	74	136	3,875
	20-24	36	154	50	119	180	5,552
	25-29	36	168	57	133	193	6,052
	30-34	36	147	52	112	172	5,302
	35-39	36	134	43	101	146	4,809
	40-44	36	124	49	82	144	4,449
	45-49	36	120	54	74	155	4,322
	50-54	36	117	37	82	142	4,214
	54-59	36	103	26	91	115	3,703

f

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
El Prat de Llobregat	60-64	36	78	23	66	86	2,803
	65-69	36	74	27	56	89	2,679
	70-74	36	58	23	44	75	2,101
	75-79	36	43	19	30	55	1,556
	80-84	36	25	10	17	33	887
	85-90	36	14	7	9	17	500
	+90	36	6	5	3	7	218
	Total	36	1,717	411	1,405	1,902	61,818
Sant Adrià del Besòs	0-4	23	73	38	50	88	1,678
	5-9	23	67	36	50	79	1,537
	10-14	23	72	38	50	90	1,655
	15-19	23	86	41	56	102	1,975
	20-24	23	112	48	85	146	2,586
	25-29	23	132	74	99	143	3,024
	30-34	23	116	64	89	138	2,674
	35-39	23	109	49	74	130	2,507
	40-44	23	95	43	66	122	2,193
	45-49	23	76	33	58	100	1,749
	50-54	23	80	38	63	101	1,847
	54-59	23	87	39	75	98	1,996
	60-64	23	70	26	59	85	1,612
	65-69	23	73	29	54	94	1,676
	70-74	23	57	22	45	75	1,301
	75-79	23	42	18	34	53	977
	80-84	23	25	12	18	33	577
	85-90	23	11	7	7	15	254
	+90	23	5	3	3	7	121
	Total	23	1,389	538	1,195	1,638	31,939

g

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Sant Feliu de Llobregat	0-4	29	70	52	41	81	2,029
	5-9	29	60	33	40	71	1,737
	10-14	29	62	29	41	75	1,787
	15-19	29	77	33	52	102	2,241
	20-24	29	117	31	84	152	3,384
	25-29	29	151	71	109	165	4,381
	30-34	29	135	89	81	140	3,921
	35-39	29	111	64	67	116	3,207
	40-44	29	96	48	59	114	2,778
	45-49	29	90	43	62	116	2,623
	50-54	29	98	33	76	122	2,843
	54-59	29	86	38	60	111	2,488
	60-64	29	60	23	42	74	1,747
	65-69	29	56	25	42	70	1,628
	70-74	29	40	18	27	54	1,174
	75-79	29	33	16	24	41	957
	80-84	29	20	11	15	27	593
	85-90	29	12	8	7	17	360
	+90	29	6	5	2	7	164
	Total	29	1,381	452	1,066	1,706	40,042
Sant Joan Despí	0-4	20	84	71	50	74	1,672
	5-9	20	75	67	49	60	1,490
	10-14	20	69	50	46	65	1,372
	15-19	20	80	47	58	90	1,606
	20-24	20	109	50	87	126	2,176
	25-29	20	139	45	117	159	2,784
	30-34	20	138	73	94	136	2,766
	35-39	20	127	99	90	108	2,541

h

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Sant Joan Despí	40-44	20	113	83	74	106	2,256
	45-49	20	95	61	62	104	1,897
	50-54	20	96	47	74	119	1,916
	54-59	20	87	42	59	108	1,744
	60-64	20	63	30	46	83	1,266
	65-69	20	56	24	47	73	1,116
	70-74	20	43	20	38	59	866
	75-79	20	33	15	30	41	650
	80-84	20	19	10	14	25	371
	85-90	20	9	5	7	12	187
	+90	20	5	3	2	8	96
	Total	20	1,439	662	1,092	1,537	28,772
Sant Just Desvern	0-4	8	96	29	90	111	767
	5-9	8	91	42	62	104	725
	10-14	8	92	41	75	113	737
	15-19	8	100	46	65	123	801
	20-24	8	124	55	81	147	992
	25-29	8	142	50	120	150	1,134
	30-34	8	134	48	112	160	1,072
	35-39	8	142	44	119	164	1,133
	40-44	8	128	46	112	139	1,025
	45-49	8	130	57	106	163	1,043
	50-54	8	127	62	75	160	1,014
	54-59	8	102	46	60	133	812
	60-64	8	78	42	42	102	627
	65-69	8	74	42	53	95	595
	70-74	8	62	34	58	84	498
	75-79	8	50	28	40	66	403

Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Sant Just Desvern	80-84	8	29	17	22	38	232
	85-90	8	18	16	10	21	143
	+90	8	21	16	4	21	117
	Total	8	1,734	637	1,252	1,999	13,870
Santa Coloma de Gramenet	0-4	99	49	17	38	60	4,811
	5-9	99	42	14	32	52	4,207
	10-14	99	48	18	37	59	4,795
	15-19	99	64	26	46	79	6,290
	20-24	99	102	39	74	120	10,104
	25-29	99	119	37	91	143	11,737
	30-34	99	99	28	78	119	9,756
	35-39	99	80	25	60	95	7,944
	40-44	99	69	28	47	85	6,822
	45-49	99	68	30	47	80	6,761
	50-54	99	82	34	55	106	8,093
	54-59	99	84	32	59	105	8,281
	60-64	99	67	22	50	86	6,647
	65-69	99	59	20	46	74	5,851
	70-74	99	43	14	33	52	4,282
	75-79	99	32	11	23	40	3,181
	80-84	99	19	8	13	23	1,911
	85-90	99	10	7	6	13	1,035
	+90	99	5	5	2	6	484
	Total	99	1,141	344	874	1,346	112,992
Tabla Total	0-4	2203	45	28	29	53	99,162
	5-9	2203	42	25	26	50	92,097
	10-14	2203	44	27	26	54	97,400
	15-19	2203	55	35	32	67	120,347

j

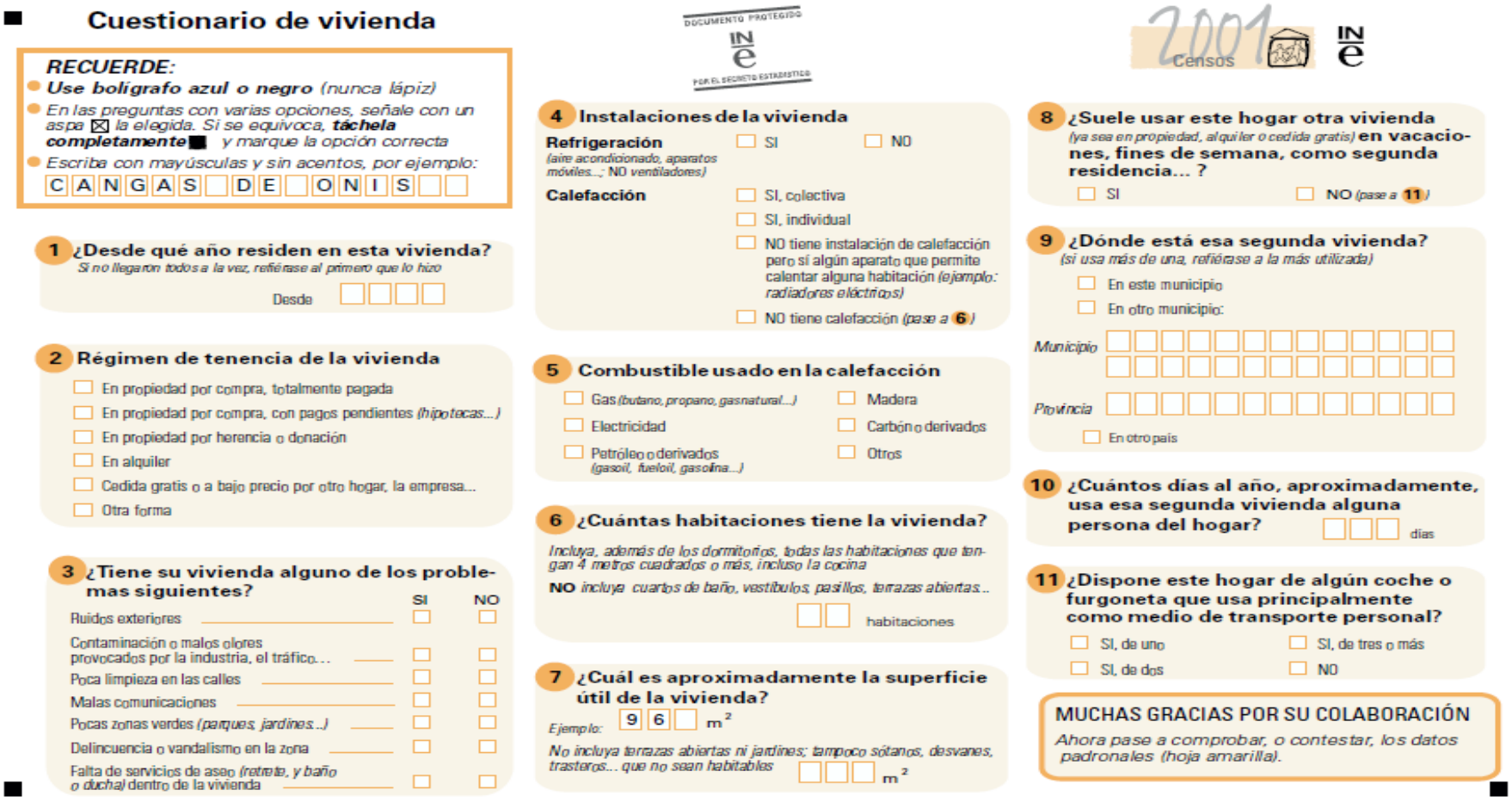
Tabla.A1 *(continuación)*

	Área Metropolitana						
	Edad	Número de áreas censales	Media	Std deviation	Percentile 25	Percentile 75	Población Total
Tabla Total	20-24	2203	80	45	50	96	176,252
	25-29	2203	97	48	65	117	212,607
	30-34	2203	88	46	61	103	193,110
	35-39	2203	82	42	57	96	180,816
	40-44	2203	76	42	49	89	166,947
	45-49	2203	70	43	43	85	155,290
	50-54	2203	72	40	45	88	159,069
	54-59	2203	69	34	45	85	151,452
	60-64	2203	57	25	40	70	125,467
	65-69	2203	62	26	44	74	135,646
	70-74	2203	55	24	39	66	120,536
	75-79	2203	44	20	31	54	97,453
	80-84	2203	27	13	18	34	59,622
	85-90	2203	15	8	9	19	32,685
	+90	2203	7	5	4	9	15,615
	Total	2203	1,086	451	782	1,276	2,391,573

k

Encuesta

Encuesta realizada a diferentes habitantes de los diferentes lugares censales.



Cuestionario de vivienda

DOCUMENTO PROTEGIDO POR EL SECRETO ESTADÍSTICO

2001 Censos

RECUERDE:

- **Use bolígrafo azul o negro** (nunca lápiz)
- En las preguntas con varias opciones, señale con un aspa ☒ la elegida. Si se equivoca, **táchela completamente** ■ y marque la opción correcta
- Escriba con mayúsculas y sin acentos, por ejemplo:

C A N G A S D E O N I S

1 ¿Desde qué año residen en esta vivienda?

Si no llegaron todos a la vez, refiérase al primero que lo hizo

Desde ☐☐☐☐

2 Régimen de tenencia de la vivienda

- ☐ En propiedad por compra, totalmente pagada
- ☐ En propiedad por compra, con pagos pendientes *(hipotecas...)*
- ☐ En propiedad por herencia o donación
- ☐ En alquiler
- ☐ Cedida gratis o a bajo precio por otro hogar, la empresa...
- ☐ Otra forma

3 ¿Tiene su vivienda alguno de los problemas siguientes?

	SI	NO
Ruidos exteriores	☐	☐
Contaminación o malos olores provocados por la industria, el tráfico...	☐	☐
Poca limpieza en las calles	☐	☐
Malas comunicaciones	☐	☐
Pocas zonas verdes *(parques, jardines...)*	☐	☐
Delincuencia o vandalismo en la zona	☐	☐
Falta de servicios de aseo *(retrete, y baño o ducha)* dentro de la vivienda	☐	☐

4 Instalaciones de la vivienda

Refrigeración *(aire acondicionado, aparatos móviles...; NO ventiladores)*

- ☐ SI ☐ NO

Calefacción

- ☐ SI, colectiva
- ☐ SI, individual
- ☐ NO tiene instalación de calefacción pero sí algún aparato que permite calentar alguna habitación *(ejemplo: radiadores eléctricos)*
- ☐ NO tiene calefacción *(pase a 6)*

5 Combustible usado en la calefacción

- ☐ Gas *(butano, propano, gasnatural...)*
- ☐ Electricidad
- ☐ Petróleo o derivados *(gasoil, fueloil, gasolina...)*
- ☐ Madera
- ☐ Carbón o derivados
- ☐ Otros

6 ¿Cuántas habitaciones tiene la vivienda?

Incluya, además de los dormitorios, todas las habitaciones que tengan 4 metros cuadrados o más, incluso la cocina

NO *incluya cuartos de baño, vestíbulos, pasillos, terrazas abiertas...*

☐☐ habitaciones

7 ¿Cuál es aproximadamente la superficie útil de la vivienda?

Ejemplo: 9 6 ☐ m^2

No incluya terrazas abiertas ni jardines; tampoco sótanos, desvanes, trasteros... que no sean habitables ☐☐☐ m^2

8 ¿Suele usar este hogar otra vivienda *(ya sea en propiedad, alquiler o cedida gratis)* **en vacaciones, fines de semana, como segunda residencia... ?**

- ☐ SI ☐ NO *(pase a 11)*

9 ¿Dónde está esa segunda vivienda?

(si usa más de una, refiérase a la más utilizada)

- ☐ En este municipio
- ☐ En otro municipio:

Municipio ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Provincia ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

- ☐ En otro país

10 ¿Cuántos días al año, aproximadamente, usa esa segunda vivienda alguna persona del hogar? ☐☐☐ días

11 ¿Dispone este hogar de algún coche o furgoneta que usa principalmente como medio de transporte personal?

- ☐ SI, de uno ☐ SI, de tres o más
- ☐ SI, de dos ☐ NO

MUCHAS GRACIAS POR SU COLABORACIÓN

Ahora pase a comprobar, o contestar, los datos padronales (hoja amarilla).

I

Tabla.A2

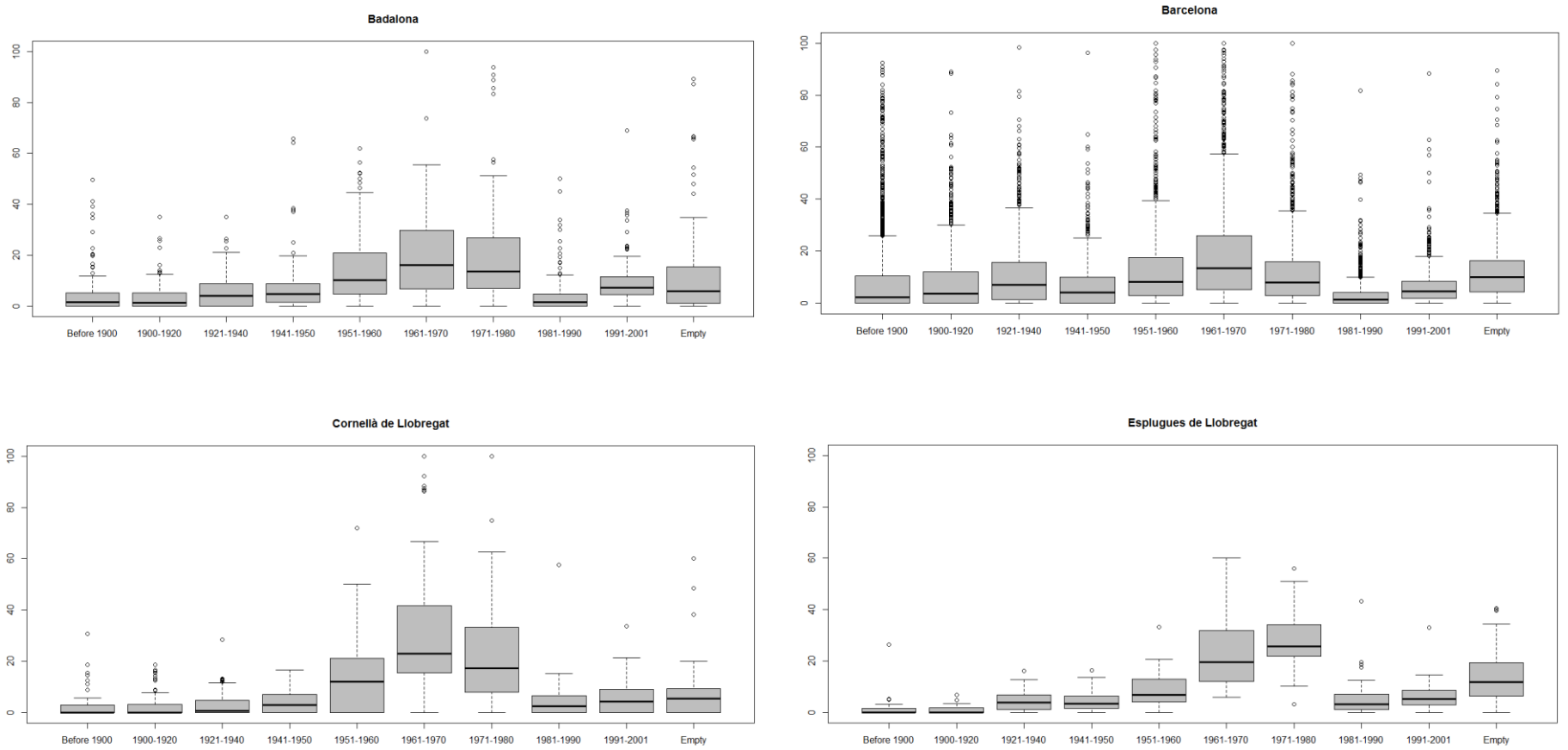
Información sobre las variables socioeconómicas

	Trabajos manuales %				Desocupados %				Bajo nivel de educaciión %			
	Mean	Std deviation	PCTL 25	PCTL 75	Mean	Std deviation	PCTL 25	PCTL 75	Mean	Std deviation	PCTL 25	PCTL 75
Badalona	66.2	14.9	60.3	76.1	16.6	7.2	12.5	18.8	17.9	8.7	12.6	23.3
Barcelona	41.5	16	29	52.3	12.6	4.6	9.6	14.5	9.8	6.7	5.4	12.2
Cornellà de Llobregat	64.6	9.8	58.1	72.95	14.1	4.2	11.1	15.5	13.2	6	9.3	16.5
Esplugues de Llobregat	50.9	16.1	42.9	59.1	11.6	6.2	8.3	12.8	9.2	6	5.6	11.2
L'Hospitalet de Llobregat	64.1	8.1	60.7	69.7	14.2	3.6	12.1	16.3	14	5.2	10.3	16.9
Montcada i Reixach	62.9	9.4	57	69.2	12.3	3.5	9.8	14.3	16.2	6	13.2	18
El Prat de Llobregat	66	10.8	60	70	15.2	7.1	11.4	16.2	17.4	9.5	12.1	20.1
Sant Adrià del Besòs	70.1	11.8	61.5	80.8	21.7	12.3	13.7	21.6	23.7	14.4	13.4	26.5
Sant Feliu de Llobregat	58.6	14.1	48.9	71.2	11.5	3.3	8.7	13.4	13.2	5.8	9.8	17.3
Sant Joan Despí	56.9	15.6	47.2	69.4	11.3	3.5	9	14.2	12	6.5	7.5	13.7
Sant Just Desvern	34.7	9.5	28.5	37.3	8.7	1.4	7.9	9.4	7	3.3	5	7.3
Santa Coloma de Gramenet	70.7	6.7	66.4	75.4	14.7	4.2	11.8	17.5	14.5	5.2	11.1	17.3
Total	48.9	1.8	34	65	13.3	5.1	10.1	15.4	11.5	7.4	6.3	15

m

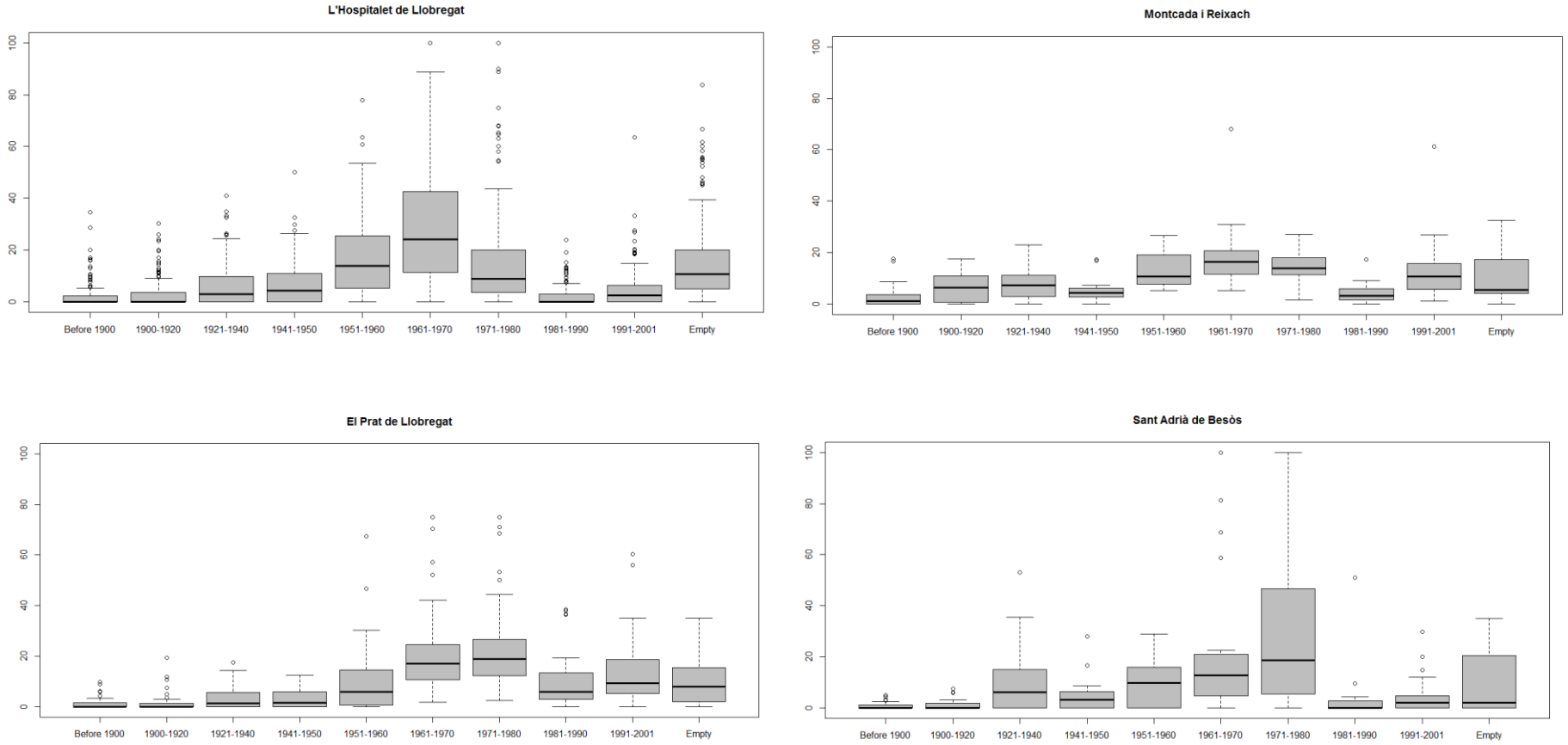
Figura.A1

Boxplots del porcentaje de construcciones de viviendas en diferentes intervalos de tiempo y porcentaje de viviendas vacías.



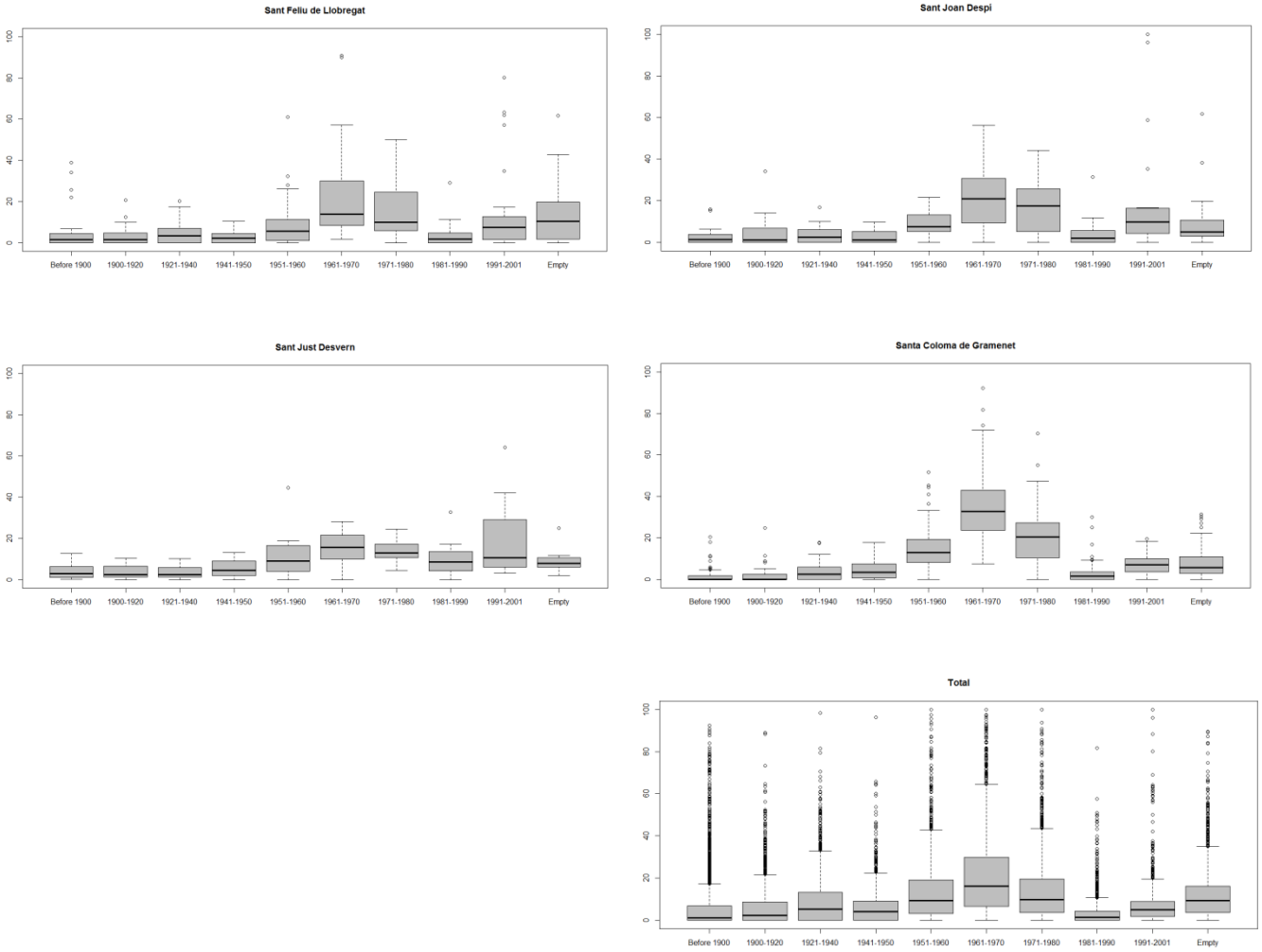
n

Figura.A1 *(continuación)*



O

Figura.A1 *(continuación)*



p

Bibliografía

[Abellan et al 2008] Abellan, J., (2008) Statistical concepts and tools for exposure and risk assessment [diapositivas]. Centre de REcerca en Epidemiologia Mediambiental (CREAL), 109 diapositivas.

[Barceló et al 2008] Barceló, M. A. et al, (2008): Métodos para la suavización de indicadores de mortalidad: aplicación al análisis de desigualdades en mortalidad en ciudades del Estado español (Proyecto MEDEA). *Gaceta Saniaria,* 22(6): 596-608.

[Barceló et al 2009] Barceló, M. A., Saez, M., Saurina, C. (2009): Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. *Science of the Total Environment,* 407(21): 5501-5523.

[Brown et al 2003] Browne, W. J., Lawson, J. B., Vidal, C. L., (2003): Disease Mapping with WinBugs and MLwiN. Ed: WILEY. Chapter 6: Relative risk estimation, 115- 139.

[Clayton et al 1987] Clayton, D., Kaldor, J., (1987): Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping. *Biometrics*, 43(3): 671-678.

[Dadvand et al 2012] Payam, D. et al, (2012): Green space, health inequality and pregnancy. *Environment International,* 40(6): 110-115.

[Getis et al 1992] Getis, A., Ord, J. K., (1992): Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 27(4): 286-306.

[Gómez-Rubio et al 2007] Gómez-Rubio, V., (2007) Analysing Spatial Data in R: Worked examples: (Bayesian) disease mapping II [diapositivas]. *Department of Epidemiology and Public Heath Imperial College London*, 16 diapositivas.

[Haining et al 2007] Haining, R, et al, 2007: Bayesian modeling of environmental: example using a small area ecological study of coronary heart disease mortality in relation to modeled outdoor nitrogen oxide levels. *Stochastic Environmental Research and Risk Assessment*, 21(5): 501-509.

[Maas et al 2006] Mass, J., et al, 2006: Green space, urbanity, and health: how strong is the relation?. *Journal of Epidemiology and Community Health*, 60(7): 587-592.

[Mitchell et al 2007] Mitchell, R., Popham, F. (2007): Greenspace, urbanity and health: relationships in England. *Journal of Epidemiology and Community Health*, 61(8): 681-683.

[Mitchell et al 2008] Mitchell, R., Popham, F. (2008): Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet*, 372(9650): 1655–1660.

[Roos et al 2011] Roos M., Held L., (2011): Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2): 259-278.

[Rue et al 2009] Rue, H., Martino, S., Chopin, N.; Approximate Bayesian Inference for Latent Gaussian Models Using Integrates Nested Laplace Approximation (with discussion). Journal of the Royal Statistical Society, Series B, 71: 319-392

[Sánchez et al 2000] Sánchez, E., et al, (2000): Comparación del NDVI con el PVI y el SAVI como Indicadores para la Asignación de Modelos de Combustible para la Estimación del Riesgo de Incendios de Andalucía. *Congreso: Tecnologías geográficas para el desarrollo sostenible. IX Congreso del Grupo de Métodos Cuantitativos, Sistemas de Información Geográfica y Teledetección. Resúmenes de presentaciones*, 164-174.

[Schrödle et al 2010] Schrödle, B., Held, L., (2010): A primer on disease mapping and ecological regression using INLA. *Computatitonal Statistics*, 26(2): 241-258.

[Shekhar et al 2008] Shekhar, S., Xiong H., (2008): Encyclopedia of CIS. Ed: SPRINGER. 923-927.

[Taylor et al 2012] Taylor, B. M., Diggle, P. J., (2012): INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes. http://arxiv.org/pdf/1202.1738v2.pdf [online: last accessed 12 April 2012].

[Thomas et al 2004] Thomas, A., Best, N., Lunn, D., Richard, A., Spiegelhalter, D. (2004): GeoBUGS User Manual. BUGS (Version 1.2).

[van den Berg et al 2010] van den Berg, A. E. et al, (2010) Green space as a buffer between stressful life events and health. *Social Science & Medicine*, 70(8): 1203-1210.

[Wakefield 2006] Wakefield J., (2006): Disease mapping and spatial regression with count data. *Biostatistics*, 8(2): 158–183.

[Waldhoer et al 2008] Waldhoer, T., Wald, M., Heinzl, H., 2008: Analysis of the spatial distribution of infant mortality by cause of death in Austria in 1984 to 2006. International Journal of Health Geographics, 7(21): 1-9.

Webs

http://www.r-inla.org/
http://earthobservatory.nasa.gov/Features/MeasuringVegetation/
http://www.r-project.org/